IMPLEMENTACIÓN DE UN MODELO DE ANÁLISIS DE SENTIMIENTOS CON RESPECTO A LA JEP BASADO EN MINERÍA DE DATOS EN TWITTER.

ERIKA PAOLA PAEZ GUARNIZO - 625659
ANDRÉS FELIPE MONROY - 625683

ASESOR
ROGER GUZMÁN
M. SC.(C) INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

UNIVERSIDAD CATÓLICA DE COLOMBIA
FACULTAD DE INGENIERÍA
PROGRAMA DE INGENIERÍA DE SISTEMAS
MODALIDAD TRABAJO DE INVESTIGACIÓN TECNOLÓGICA
BOGOTÁ D.C.
2020

IMPLEMENTACIÓN DE UN MODELO DE ANÁLISIS DE SENTIMIENTOS CON RESPECTO A LA JEP BASADO EN MINERÍA DE DATOS EN TWITTER.

ERIKA PAOLA PAEZ GUARNIZO - 625659
ANDRES FELIPE MONROY - 625683

ESTE TRABAJO DE GRADO ES PRESENTADO COMO REQUISITO PARA OPTAR AL TÍTULO DE: INGENIERO DE SISTEMAS

ASESOR:
ROGER ENRIQUE GUZMÁN AVENDAÑO
M. SC (C). INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

ALTERNATIVA:
TRABAJO DE INVESTIGACIÓN TECNOLÓGICA
GRUPO DE INVESTIGACIÓN:
GISIC
SEMILLERO DE INVESTIGACIÓN:
MAILAB

UNIVERSIDAD CATÓLICA DE COLOMBIA
FACULTAD DE INGENIERÍA
PROGRAMA DE INGENIERÍA DE SISTEMAS
MODALIDAD TRABAJO DE INVESTIGACIÓN TECNOLÓGICA
BOGOTÁ D.C.
2020

Nota de aceptación

Aprobado por el comité de grado en cumplimiento de los requisitos exigidos por la Facultad de Ingeniería y la Universidad Católica de Colombia para optar al título de ingeniero de sistemas.

_____

Juan Carlos Barrero

Jurado 2

_____

Roger Enrique Guzmán Avendaño, Msc.

Asesor

BOGOTÁ D.C. JUNIO 13 DE 2020

DEDICATORIA

Dedicamos este proyecto a Dios, nuestros padres, hermanos y abuelas. A Dios porque siempre ha estado con nosotros, cuidándonos y dándonos fortaleza; a nuestros padres, quienes a lo largo de nuestras vidas han velado por nuestro bienestar y educación siendo nuestro apoyo en todo momento, depositando en nosotros su entera confianza en cada reto a lo largo de este camino que se nos presentaba, sin dudar en algún momento de nuestras capacidades, fortalezas e inteligencia.

AGRADECIMIENTOS

Expresamos nuestro profundo agradecimiento a Dios por darnos fortaleza y constancia para lograr una meta y un objetivo muy importante en nuestras vidas. A nuestros padres por ser nuestra motivación y una gran fortaleza para todos los días llegar a ser unas mejores personas y profesionales, también por su apoyo, comprensión y ayuda por lograr hacer uno de nuestros sueños realidad, que es ser unos profesionales, ya que sin ellos no hubiéramos logrado nuestra meta.

De manera muy especial agradecemos a nuestro director el Ingeniero Roger Enrique Guzmán Avendaño, ya que, sin su apoyo, ayuda, conocimientos y capacidades no hubiéramos logrado desarrollar este proyecto, el cual se logró con mucho esfuerzo y dedicación.

Finalmente, a la Universidad Católica De Colombia, a los docentes que nos impartieron clases a lo largo de este proceso y a la dirección del programa.

TABLA DE CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

ECUACIONES

Anexos

RESUMEN

Este documento presenta un experimento de métodos basados en aprendizaje de máquina y minería de datos con el fin que mediante el tema de la Jurisdicción Especial para la Paz (JEP), implementar tres algoritmos los cuales fueron seleccionados mediante el estado del arte, en los que se realizó una comparación con las características de unigramas y bigramas de esta forma generando un modelo de análisis de sentimientos de las personas (positivo, negativo y neutro) y la opinión en este tema político, donde según estadísticas, las personas no se encuentran de acuerdo con los rubros asignados y los gastos que la JEP ha llegado a generar. Por parte de los distintos partidos políticos se han implementado diferentes acuerdos para una paz duradera, pero las personas mediante la red social Twitter han llegado a manifestar sus distintas opiniones.

Los métodos usados en este trabajo se encuentran conformados por los siguientes algoritmos: Naive Bayes, Random Forest y Máquinas de Soporte Vectorial, los cuales permiten clasificar y conocer el sentimiento de los usuarios en Twitter con respecto al tema de la JEP.

La finalidad del presente trabajo es diseñar e implementar técnicas basadas en minería de datos para analizar la posición política de los usuarios con respecto a la JEP y conocer el algoritmo con mejor desempeño para realizar análisis de sentimientos.

Se propone una metodología que cuenta con seis (6) etapas, las cuales son: construcción del conjunto de datos, procesamiento del lenguaje natural sobre los tweets, seguido de la extracción de características, el muestreo de los datos por medio de validación cruzada en el caso de Máquinas de Soporte vectorial, por otra parte a los algoritmos de Naive Bayes y Random Forest la segmentación a los datos se realiza en los porcentajes de muestreo 70-75-80% y en los porcentajes de testeo con un valor de 30-25-20%, después se realiza el entrenamiento y clasificación para cada algoritmo, por último, la evaluación del desempeño de las técnicas implementadas. En la última etapa se evidencia el mejor método de clasificación de texto, en el cual su resultado se dio en el algoritmo de Random Forest con métricas de precisión con un valor de 74,56%, recall 70,15% y F1-score con un resultado de 68,10%

Palabras clave: Análisis de Sentimientos, JEP, Machine Learning, Máquinas de Soporte Vectorial, Minería de datos, Naive Bayes, Random Forest, Twitter.

ABSTRACT

This document presents an experiment in methods based on machine learning and data mining in order to find out, how the country is polarized, regarding to the special jurisdiction for peace (JEP), implement three algorithms which were selected using the state of the art, in which a comparison was made with the characteristics of unigrams and bigrams in this way, generating a model of analysis of people's feelings (positive, negative and neutral) and opinion on this political issue, where according to statistics, people do not agree with the items assigned and the expenses that the JEP has come to generate. Different agreements for a lasting peace have been implemented by the different political parties, but people through the social network Twitter have come to express their different opinions.

The methods used in this work, are conformed by the following algorithms: Naive Bayes, Random Forest and Support Vector Machine, the ones we choose and know the sentiment of the users on Twitter regarding the issue of the JEP.

The purpose of this work is to design and implement techniques based on data mining to analyze the political position of users with respect to the JEP and to know the algorithm with the best performance to perform sentiment analysis.

It proposes a methodology with six (6) stages, which are construction of the data set, processing of the natural language on the tweets, followed by the extraction of characteristics, the sampling of the data by means of crossed validation in the case of Machines of Vectorial Support, on the other hand to the algorithms of Naive Bayes and Random Forest the segmentation to the data is made in the percentages of sampling 70-75-80% and in the percentages of testing with a value of 30-25-20%, later the training and classification for each algorithm is made, finally, the evaluation of the performance of the implemented techniques. In the last stage, the best text classification method is demonstrated, in which the result was given in the Random Forest algorithm with precision metrics with a value of 74.56%, remember 70.15% and F1 score with a result of 68.10%.

Keywords: Data mining, Twitter, JEP, Sentiment Analysis, Machine Learning, Naive Bayes, Random Forest, Suppor Vector Machine.

INTRODUCCIÓN

A lo largo de los últimos años las redes sociales se han convertido en una parte esencial para lograr expresarse, las redes sociales son importantes en cada una de las áreas que se desarrollen, a su vez se debe tener mucho cuidado con las mismas, porque puede hacer daño a personas, productos o compañías y debemos aprender a usarlas de manera responsable[1], debido a esto se han convertido en herramientas que permite conocer diferentes opiniones a lo largo del mundo, las opiniones son esenciales cuando queremos conocer la percepción que tienen las personas acerca de un tema.

Una de las plataformas en la que los usuarios pueden dar una opinión respecto a temas políticos, culturales y sociales es Twitter, puesto que no solamente las personas pueden hacer uso de ella, sino que también es muy usada para los temas políticos en la que se realizan tendencias para las candidaturas presidenciales, un ejemplo de esto fue la reelección del presidente Barack Obama[2]. En Colombia estas tendencias se pueden implementar para lograr conocer las distintas opiniones frente a uno de los temas más emblemáticos en el país que es el posconflicto, las decisiones acerca de este tema centrándonos especialmente en la Jurisdicción Especial para la Paz (JEP) son muy divididas, razón por la cual se quiere conocer el sentimiento de los usuarios que dan una opinión en esta red social, utilizando minería de datos y aprendizaje de máquina con el propósito de conocer la polarización de las personas que han realizado trinos en español con la palabra JEP en los últimos dos años.

La minería de datos nos permite identificar patrones en conjuntos de datos grandes, una de sus características es ser predictiva teniendo la posibilidad de indicar que es lo que pasara utilizando estadísticas y probabilidades de información que está oculta en datos almacenados[3]. Para realizar predicciones puede utilizar el aprendizaje de máquinas o aprendizaje automático que es un subcampo de las ciencias de la computación y la inteligencia artificial el cual tiene como objetivo “desarrollar técnicas que permitan a las computadoras aprender en el sentido de que se crean programas capaces de generalizar comportamientos a partir de una información no estructurada suministrada en forma de ejemplos”[4], del aprendizaje automático se desprende el aprendizaje supervisado en el cual se agrupan los algoritmos que trabajan a partir de datos etiquetados, estos algoritmos utilizan un histórico de datos para ser entrenados

[1] http://www.utp.ac.pa [en línea] Las Redes Sociales y su Impacto en la Sociedad de Hoy < http://www.utp.ac.pa/las-redes-sociales-y-su-impacto-en-la-sociedad-de-hoy#:~:text=Marcela%20Madrid%20Guerra%2C%20las%20redes,a%20usarlas%20de%20manera%20responsable.>

[2] diarioinformacion.com [en línea] El éxito de Obama y la minería de datos <https://www.diarioinformacion.com/opinion/2012/11/17/exito-obama-mineria-datos/1315856.html>

[3] https://www.syloper.com [en línea] ¿Para qué sirve la minería de datos? – Data mining <https://www.syloper.com/blog/recursos/para-que-sirve-la-mineria-de-datos/#:~:text=La%20miner%C3%ADa%20de%20datos%20(data,datos%20en%20un%20contexto%20espec%C3%ADfico. >

[4] Peter Lang. Studien zur romanischen sprachwissenshcaft und interkulturellen kommunikation – Internationaler verlag der Wissencharften – Chapter 2 Page 101

con el propósito de predecir un valor de salida. Estos algoritmos se pueden basar en modelos probabilísticos por ejemplo Naive Bayes, modelos lógicos como lo es Random Forest o modelos geométricos como los son las Maquinas de soporte vectorial, los algoritmos mencionados han obtenido los mejores resultados en el análisis de sentimientos, tarea que se centra en catalogar o clasificar los documentos (tweets en este caso) en función de la connotación positiva o negativa del lenguaje ocupado en el mismo[5]: Para lograr esto, es muy importante el procesamiento del lenguaje natural que es otro subcampo de la inteligencia artificial el cual se centra en la interacción entre las máquinas y los lenguajes humanos.

En este documento se presenta una metodología implementada a lo largo del desarrollo de la propuesta y compuesta por la construcción del conjunto de datos, el procesamiento del lenguaje natural, la extracción de características , el muestreo de los datos, el entrenamiento y clasificación para cada algoritmo y finalmente la evaluación del desempeño de las técnicas implementadas para realizar análisis de sentimientos de los usuarios de Twitter con respecto a la JEP implementando un diseño experimental de minería de datos y entrenamiento de maquina con la finalidad de conocer la tendencia que esta genera en los usuarios de Twitter en Colombia.

[5] itelligent.es [en línea] Análisis de sentimiento, ¿qué es, cómo funciona y para qué sirve? <https://itelligent.es/es/analisis-de-sentimiento/#:~:text=El%20an%C3%A1lisis%20de%20sentimientos%2C%20tambi%C3%A9n,lenguaje%20ocupado%20en%20el%20mismo.>

1. GENERALIDADES

1.1. LÍNEA DE INVESTIGACIÓN

El proyecto pertenece al grupo de investigación en Software Inteligente y Convergencia Tecnológica – GISIC, semillero MAILAB. Debido a que esta línea de investigación aborda proyectos en el campo de minería de datos, en relación a la clasificación de texto se utiliza aprendizaje de máquina.

El cual compone diferentes ramas como lo es aprendizaje de máquina, que se encuentra compuesto por distintos algoritmos, el experimento que se aborda en este proyecto utiliza algoritmos de aprendizaje supervisado, ya que las técnicas de estos mismos logran un gran desempeño en el momento de la clasificación de texto.

1.2 PLANTEAMIENTO DEL PROBLEMA

1.2.1 Descripción del problema

La Jurisdicción Especial para la Paz (JEP), es el componente de justicia del Sistema integral de verdad, justicia, reparación y no repetición (SIVJRNR), creado en el acuerdo final para la terminación del conflicto y la construcción de una paz estable y duradera. La JEP, como mecanismo de justicia transicional, tiene la tarea de investigar, esclarecer, juzgar y sancionar los más graves crímenes ocurridos en Colombia durante más de 50 años de conflicto armado, y hasta el 1 de diciembre. de 2016 [6]. Para el 2020 la JEP se ha convertido en uno de los temas más polémicos en el tema política en un país como Colombia ya que según el periódico El Mundo de España, la corte constitucional critica "La creación de la JEP mediante normas oscuras e incoherentes, sin delimitar competencia; la elección de los magistrados de la JEP por parte de extranjeros; la posesión de los magistrados de la JEP antes de expedir las normas procesales y estatutarias; la falta de prevención en materia presupuestal para cumplir lo pactado (y hay incumplimientos), entre otros puntos deleznables."[7].

Al ser la Jurisdicción Especial para la Paz un tema tan emblemático como ya se mencionó, las redes sociales suelen ser un medio bastante común para que las personas indiquen su punto de vista acerca de este tema. En los últimos años una de la redes sociales más usadas en Colombia es Twitter según el diario el tiempo, el 11% de la población urbana en Colombia usa esta red social (Redacción Redes Sociales, 2015) [8], twitter se ha convertido en una de las plataformas más importantes frente a las opiniones políticas, ambientales, sociales y económicas, un ejemplo de ello es cuando se realizan las elecciones populares como lo son las presidenciales, alcaldías, gobernaciones, entre otros, en el territorio nacional [9], también se discuten problemáticas de carácter moral o de interés ciudadano tales como el posconflicto en Colombia, una de estas es cómo se involucra a los grupos armados a la JEP y las decisiones tomadas acerca de los acuerdos de paz.

La minería de datos se usa para realizar predicciones en el comportamiento de los de usuarios de Twitter aplicado en el contexto de redes sociales, económicos y políticos[10]. Estados unidos es uno de los países que ha

[6] ¿Qué es la jurisdicción especial para la paz? [en línea] <https://www.jep.gov.co/Infografas/conozcalajep.pdf>
[7] Acuerdo de paz: errores en serie [en línea] < https://www.elmundo.com/noticia/Acuerdo-de-pazerrores-en-serie/377466>
[8] eltiempo.com [en línea] El 11% de la población urbana en Colombia usa Twitter Disponible en internet <https://www.eltiempo.com/archivo/documento/CMS-12406882>.

[9] eltiempo.com [en línea] Lo mejor del 2018 en Twitter: las cuentas y los hashtags más populares <https://www.eltiempo.com/tecnosfera/novedades-tecnologia/las-cuentas-y-las-etiquetas-mas-populares-en-twitter-durante-2018-301996>.

[10] ars-uns.blogspot.com [en línea] Análisis de redes sociales <http://ars-uns.blogspot.com/2015/03/la-mineria-de-datos-de-twitter.html>

implementado la minería de datos para ejecutar campañas políticas en las elecciones presidenciales y de esta manera se llegar a conocer los análisis de tendencias, un claro ejemplo de ello fue en la reelección presidencial de Barack Obama [11] y en la elección presidencial de Donald Trump [12]. En Colombia se han realizado diferentes estudios con minería de datos para predecir cuál es el candidato ganador a las elecciones[13], pero no se ha ejecutado un análisis frente al sentimiento que genera en los usuarios de twitter mediante temas emblemáticos en el país como es el de la JEP.

Para conocer el sentimiento de los usuarios de las redes sociales frente a un tema en específico se puede hacer uso de la minería de datos, muchas veces estableciendo segmentaciones o divisiones de opinión, puesto que según el espectador "el 47% de colombianos tienen una opinión favorable de la JEP" [14]. Pero muchas personas no se encuentran de acuerdo con los gastos generados por la JEP, un ejemplo de esto según la FM "A algunas personas les ha generado suspicacia que dentro del rubro de sueldos y salarios (47.292 millones de pesos) se incluyeran bonificaciones por $16.682 millones y gastos de representación por $2.582 millones. También que dentro del rubro de prestaciones sociales (14.897 millones de pesos) se incluyera una prima de navidad por 2.670 millones de pesos."[15] Además la gente no se encuentra a favor con respecto a los beneficios hacia los ex miembros de los grupos armados y también la manera en que la JEP está siendo ejecutada. Estas razones tanto económicas como sociales son las causantes de que las diferentes opiniones acerca de este tema difieran, pero esto también depende de la convicción política del usuario que realice el tweet, el cual puede ser de izquierda, derecha o centro; estas opiniones pueden ayudar en el momento de tomar decisiones en el país. Uno de los mecanismos de participación ciudadana es el plebiscito en el que la inversión según el Portafolio va "Alrededor de los 350.000 millones de pesos"[16], como se observa en la Figura

[11] diarioinformacion.com [en línea] El éxito de Obama y la minería de datos <https://www.diarioinformacion.com/opinion/2012/11/17/exito-obama-mineria-datos/1315856.html>

[12] bbc.com [en línea] Elecciones en Estados Unidos ¿Fue facebook la clave para el triunfo de Donald Trump? <https://www.bbc.com/mundo/noticias-internacional-37946548>

[13] javeriana.edu.co [en línea] <https://repository.javeriana.edu.co/bitstream/handle/10554/20516/CaicedoOrtizLuisEduardo2016.pdf?sequence=1&isAllowed=y>

[14] elespectador.com [en línea] El 47% de los colombianos tiene una opinion favorable de la jep <https://www.elespectador.com/noticias/politica/47-de-colombianos-tienen-una-opinion-favorable-de-la-jep-gallup-poll-articulo-861085>

[15] lafm.com [en línea] Las polémicas cuentas de la jep <https://www.lafm.com.co/judicial/las-polemicas-cuentas-de-la-jep>

[16] portafolio.com [en línea] El plebiscito por la paz cuesta $350.000 millones, ¿qué se puede hacer con ese mismo dinero? <https://www.portafolio.co/tendencias/cuanto-cuesta-el-plebiscito-por-la-paz-499066>

1, estos son los uno de los dineros presupuestales del país, usados para el funcionamiento de la JEP. Esta gran suma de dinero según el Portafolio puede ser usado en cosas para el país como lo son “Las acreencias de EPS a la red pública hospitalaria del Valle, La reubicación del mercado de Cartagena y algunos de los proyectos, dentro del plan de desembotellamiento de las entradas y salidas de Bogotá se encuentran en torno a esa cifra” [17] .

Figura 1 Recursos en el ministerio de hacienda en el fondo Colombia en paz.

Fuente: Boletín estadístico [autor] Secretaria ejecutiva Jurisdicción especial para la paz, Disponible en internet: <https://www.jep.gov.co/Sala-de-Prensa/Documents/Boletin%20Estadistico%20abril%202018%20%281%29.pdf>

Mediante el análisis de sentimientos en la red social twitter se reduciría las encuestas que se ejecutan tradicionalmente antes de una decisión política importante en el país, además este proceso se realizaría con mayor eficacia y a su vez Colombia implementaría la tecnología con temas basados en algoritmos, las redes sociales, procesamiento de lenguaje natural y medidas de desempeño.

[17] portafolio.com [en línea] El plebiscito por la paz cuesta $350.000 millones, ¿qué se puede hacer con ese mismo dinero? <https://www.portafolio.co/tendencias/cuanto-cuesta-el-plebiscito-por-la-paz-499066>

1.2.2 Formulación del problema

Mediante el problema planteado anteriormente, la pregunta de investigación es la siguiente:

¿Cómo identificar el sentimiento positivo, negativo o neutro de los usuarios de la red social twitter con respecto a la JEP usando minería datos?

1.3 OBJETIVOS

1.3.1 Objetivo general

Implementar una técnica basada en minería de datos con el fin de realizar análisis de sentimientos con respecto a la Jurisdicción Especial para la Paz en la red social twitter.

1.3.2 Objetivos específicos

- Construir un conjunto de datos de la red social twitter con referencia al tema de la JEP, implementado herramientas que puedan conectarse al API de twitter para descargar su data.
- Diseñar una estrategia de minería de datos para analizar los sentimientos de los tweets respecto a la JEP, realizando una investigación con la cual se busca ordenar y ejecutar una serie de pasos para realizar este proceso de manera óptima.
- Implementar un modelo de minería de datos para clasificar los tweets en base a los sentimientos de los usuarios para conocer su posición con respecto a la JEP, aplicando algoritmos de aprendizaje de máquina al conjunto de datos.
- Evaluar el rendimiento de la técnica basada en minería de datos utilizando la métrica de precisión, recall, F1-score y exactitud, para identificar cual fue el mejor resultado.

1.4 JUSTIFICACIÓN

El análisis de sentimientos tiene una aplicación útil en el momento de realizar monitorización en las redes sociales sobre los cometarios que los usuarios realicen sobre algún producto o un tema en particular, esto es implementado en diferentes compañías para la potencialización de productos o la mejora de los mismos. A su vez este proceso no es implementado solamente en la industria, sino que se ha efectuado en diferentes temas sociales, culturales y políticos a nivel mundial[18].

En el proceso del diseño de análisis de sentimientos se combinan distintas áreas con las técnicas de minería de datos, algoritmos de aprendizaje supervisado y procesamiento de lenguaje natural. Se realizan estos procedimientos ya que el lenguaje humano es complejo y enseñar a una maquina a analizar los diferentes matices gramaticales, variaciones culturales, jergas y faltas de ortografía es un proceso difícil. Además, este proceso no solo valora la opinión como positiva, negativa o neutra, sino también por la detección de tendencias[19].

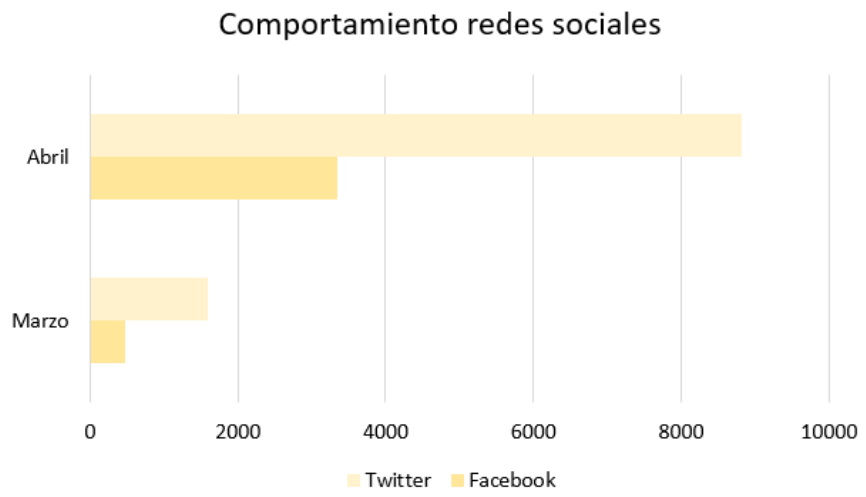
Se va a utilizar la red social Twitter, para la recolección del conjunto de datos, como se evidencia en la Figura 2 la cantidad de seguidores con la que cuenta la JEP en Twitter es mayor que en Facebook esto hará que el proceso se realice de una manera óptima y precisa. Por otra parte según MinTic en Colombia cerca de seis millones de personas usan twitter[20].

[18] Brandwatch.com [en línea] Entendiendo el análisis de sentimientos, que es y para que se usa <https://www.brandwatch.com/es/blog/analisis-de-sentimiento/>

[19] Biblogtecarios.es [en línea] el análisis de sentimiento de texto en las redes sociales <https://www.biblogtecarios.es/inmaherrero/el-analisis-de-sentimiento-de-texto-en-las-redes-sociales/>

[20] mintic.gov.co [en línea] Colombia es uno de los países con más usuarios en redes sociales en la región<https://mintic.gov.co/portal/604/w3-article-2713.html?_noredirect=1>

Figura 2 Comportamiento redes sociales



Fuente: Boletín estadístico [autor] Secretaria ejecutiva Jurisdicción especial para la paz, Disponible en internet: <https://www.jep.gov.co/Sala-de-Prensa/Documents/Boletin%20Estadistico%20abril%202018%20%281%29.pdf>

En Colombia se ha implementado el análisis de sentimientos utilizando la minería de datos para lograr conocer qué candidato puede ser elegido para la alcaldía o a la presidencia[21], pero no existe una implementación con respecto a la JEP. Las objeciones del presidente Duque[22] al tema han generado polarización en el país, esto hace que surjan diferentes sentimientos y opiniones. Con ello se desea conocer el sentimiento generado por las decisiones que se tomaran por parte de gobierno para la incorporación de las personas en el proceso de paz[23]. Otra razón por la que se quiere implementar un análisis de sentimientos con respecto a la JEP puesto que es un tema nuevo en el país, el cual ha hecho que surjan diferentes posiciones y pensamientos en las personas. Además, se invierte bastante dinero en encuestas[24] las cuales no abarcan el total de la población y las cuales no tienden a dar resultados óptimos. Con esto el análisis de sentimientos busca tener en cuenta las tendencias que está tomando el país con respecto a las decisiones que se

[21] repository.javeriana.edu.co [en línea] análisis del sentimiento político mediante la aplicación de herramientas de minería de datos a través del uso de redes sociales <https://repository.javeriana.edu.co/bitstream/handle/10554/20516/CaicedoOrtizLuisEduardo2016.pdf?sequence=1&isAllowed=y>

[22] elespectador.com [En linea]Objeciones del presidente Duque a la JEP: más políticas que de conveniencia < https://www.elespectador.com/colombia2020/justicia/jep/objeciones-del-presidente-duque-la-jep-mas-politicas-que-de-conveniencia-articulo-857740>

[23] las2orillas.co/ [en línea] Aumenta la polarización política tras la decisión de la JEP sobre Santrich <https://www.las2orillas.co/polarizacion-politica-decision-jep/>

[24] elespectador.com [en línea] 47 % de colombianos tienen una opinión favorable de la JEP: Gallup Poll <https://www.elespectador.com/noticias/politica/47-de-colombianos-tienen-una-opinion-favorable-de-la-jep-gallup-poll-articulo-861085>

toman acerca de la Jurisdicción Especial para la Paz, en lugar de implementar encuestas.

1.5 Delimitaciones

1.5.1 Limitaciones

Para el experimento a realizar solo se tendrá en cuenta el tema político de la JEP en Colombia y la clasificación de los sentimientos en tres categorías (positivos, neutros y negativos), solo se usará la red social twitter, los tweets con una antigüedad no mayor a dos años y la cantidad de tweets dependerá de las solicitudes permitidas por el API de twitter en su capa gratuita. También se debe tener en cuenta una limitante en el desarrollo del proyecto, que se puede dar en el momento de realizar procesamiento de datos mediante los equipos que se implementaran, esto puede limitar el volumen de datos a procesar en la herramienta establecida.

1.5.2 Alcances

Se desarrollará un experimento para la clasificación de sentimientos en la red social twitter sobre el tema de Jurisdicción especial para la paz (JEP), donde se tendrá en cuenta los sentimientos positivos negativos y neutros, sentimientos que serán etiquetados en cada tweet que cumpla las características descritas en el documento con el propósito de entrenar tres algoritmos de aprendizaje de máquina y escoger cual es el mejor de los tres, dicho experimento se desarrollará en el lenguaje de programación Python 3 y R durante el periodo académico de primer semestre del 2020.

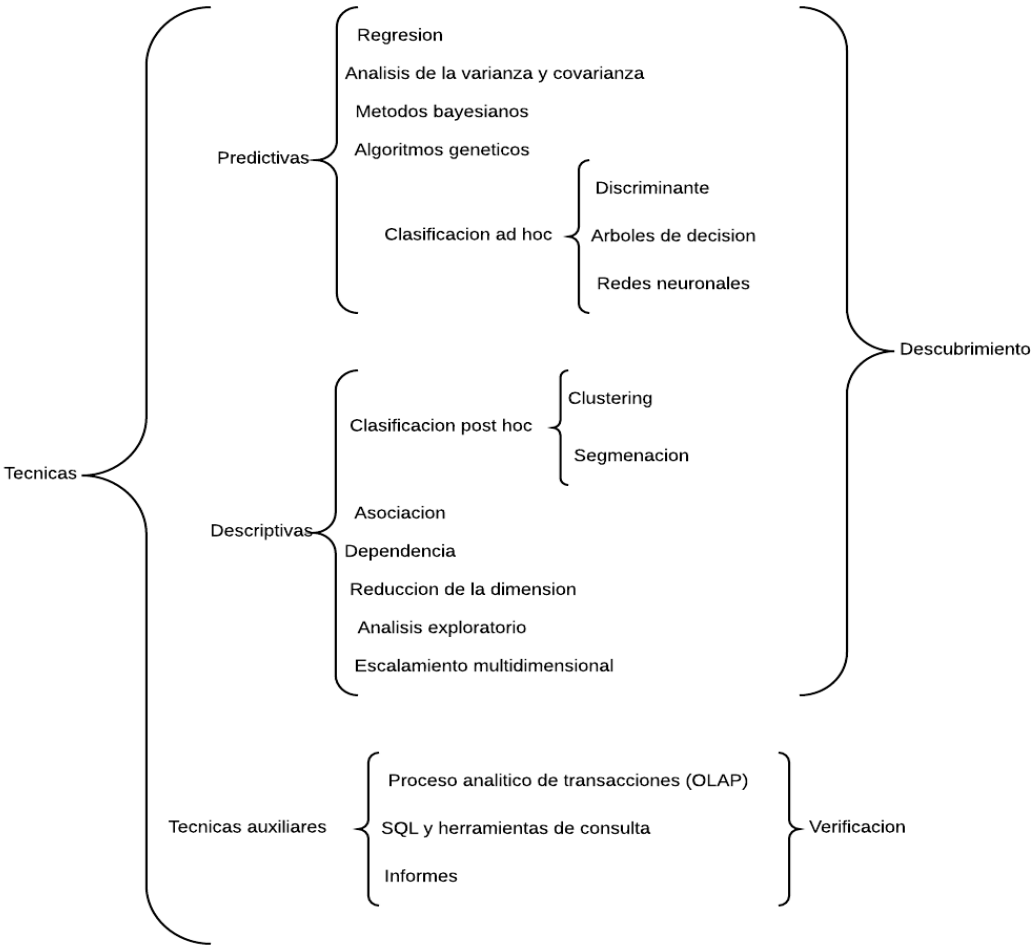
2 MARCO REFERENCIAL

A continuación, se describe el marco teórico y el marco conceptual.

2.2 MARCO TEÓRICO

En esta sección se describen los métodos para realizar el proceso de minería de datos, como se puede observar en la Figura 1 *Recursos en el ministerio de hacienda en el fondo Colombia en paz.*

Figura 3 Técnicas de minería de datos



Fuente: Minería de datos técnicas y herramientas [autor] César Pérez López Disponible en internet: <https://books.google.com.co/books/about/Miner%C3%ADa_de_datos.html?id=wz-D_8uPFCEC&redir_esc=y>

2.2.1 Minería de datos predictiva

La minería de datos predictiva consiste en realizar un método por medio de diferentes datos para lograr predecir su comportamiento en relación con una o más variables. Este se lleva a cabo mediante la búsqueda de normas de clasificación o de predicción basado en los resultados que se pueden llegar a tener en el futuro. Consiste en la extracción de información existente en los datos y su uso para la predicción de tendencias y patrones. [25]

Esta tiene como objetivo el ajuste de la predicción para una variable t, dándose este en un nuevo valor de variable de entrada x sobre la base de un conjunto de datos de entrenamiento N el cual tiene como valor $x = (x1, \dots, xn)$ y sus valores objetivos de $t = (t1, \dots, tn)$, en el cual se puede expresar la incertidumbre sobre un valor de variable objetivo. Para determinar su valor de parámetros wML el cual utiliza la media y βML para encontrar su precisión, con el cual se puede realizar nuevos valores de x. En el cual se implementa para una estimación puntual de parámetros de probabilidad[26], como se observa en la Ecuación 1

Ecuación 1 Minería de datos predictiva

$$p(x, wML, \beta ML) = N(t|y(x, wML), \beta^{-1}ML)$$

2.2.2 Regresión

Los algoritmos de regresión se usan para predecir los valores ausentes en una o más variables continuas, las cuales se pueden predecir en pérdidas o ganancias. [27]

El modelo lineal de regresión tiene como objetivo analizar los datos que pueden surgir y que estos se representan en un valor para el mínimo alcanzable de una pérdida esperada, esto tiene como nombre ruido, teniendo $y(x)$ como la elección para una función; la cual realiza la solución mediante un término mínimo, sin embargo, el conjunto de datos es representado por D en el cual se obtiene un número finito N de los puntos dados y siendo $h(x)$ la regresión de la función. Teniendo así la diferencia al cuadrado entre $y(x; D)$ y $h(x)$ la cual se expresa como la suma de dos términos, el primero llamado sesgo cuadrado en el que se representa el grado de predicción sobre todos los conjuntos de datos, el segundo término es la varianza la cual mide la solución para los datos

[25] bigdata-social.com [en línea] Análisis predictivo<http://www.bigdata-social.com/que-es-el-analisis-predictivo/ >

[26] Christopher M. Bishop. Pattern recognition and machine learning: Linear models for regression En: Information Science and Statics. Pages 45 – 46.

[27] docs.microsoft.com [en línea] Algoritmos de minería de datos<https://docs.microsoft.com/es-es/analysis-services/data-mining/data-mining-algorithms-analysis-services-data-mining?view=sql-server-2017 >

individuales que varían por su promedio.
Ecuación 2 Varianza

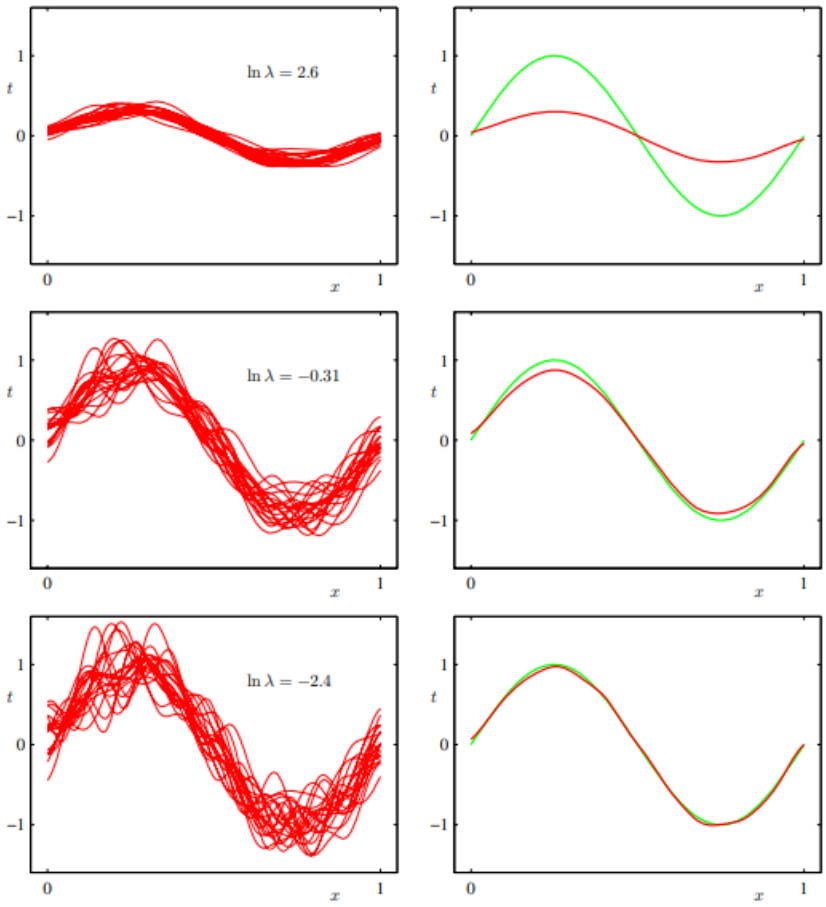
$$varianza = \int E_D[\{y(x;D) - E_D[y(x;D)]\}^2]p(x)dx$$

Ecuación 3 Ruido

$$ruido = \int \{h(x) - t\}^2 p(x,t)dxdt$$

El objetivo es minimizar la pérdida esperada, la cual tiene en cuenta la suma de varianza y ruido constante. Como se puede observar en la Figura 4 el modelo con la capacidad predictiva es que da el mejor equilibrio entre el sesgo y la varianza realizando una independencia entre estos dos, teniendo en cuenta $L = 100$ conjuntos, cada uno con $N = 25$ datos, donde la columna izquierda muestra el resultado de ajustar el modelo a los conjuntos de datos para diferentes valores Ln y la columna derecha muestra el promedio de los 100 ajustes que se ven en color rojo y en color verde se observa el conjunto de datos.[28]
Figura 4 Modelo de regresión lineal

[28] Christopher M. Bishop. Pattern recognition and machine learning: Linear models for regression En: Information Science and Statics. Contents 3. Pages 149 – 150.



Fuente: Christopher M. Bishop. Pattern recognition and machine learning: Linear models for regression En: Information Science and Statics. Contents 3. Pages 149 – 150.

2.2.3 Métodos bayesianos

El método bayesiano es un algoritmo de clasificación el cual se usa para el modelamiento de predicción y exploración realizando una cuantificación de la incertidumbre. "útil para generar rápidamente modelos de minería de datos que detectan las relaciones entre las columnas de entrada y las columnas de predicción. Puede utilizar este algoritmo para realizar la exploración inicial de los datos y, más adelante, aplicar los resultados para crear modelos de minería de datos adicionales". [29] Siendo así w un modelo de cuantificador de incertidumbre parámetro, siendo $p(w)$ la probabilidad previa de distribución y la de los datos observador $D = \{t1, \dots, tN\}$ que se expresa en una probabilidad condicional $p(D|w)$.[30]

[29] docs.microsfot.com [en línea] Naive Bayes Algorithm <https://docs.microsoft.com/es-es/analysis-services/data-mining/microsoft-naive-bayes-algorithm?view=sql-server-2017 >

[30] Christopher M. Bishop. Pattern recognition and machine learning: Linear models for regression En: Information Science and Statics. Contents 2. Pages 90 – 97.

Ecuación 4 Métodos Bayesianos

$$p(D) = \frac{p(w)p(w)}{p(D)}$$

2.2.4 Discriminante

Esta es una técnica multivariante con la cual se logra describir las diferencias entre los grupos, con los cuales se logra observar determinadas variables. Uno de los usos principales en la clasificación de los grupos son las variables preestablecidas y descriptivas.[31] Este se encuentra clasificado en dos etapas separadas, la primer etapa es la de inferencia en la cual se usa los datos de entrenamiento para un modelo $p(Ck|x)$ y la segunda es la etapa de decisiones en la que se usa probabilidades para realizar asignaciones de clase óptimas.[32]
Ecuación 5 Discriminante

$$p(x) = \frac{p(x|Ck)p(Ck)}{p(x)}$$

2.2.5 Árboles de decisión

El árbol de decisión es un algoritmo con el cual se realiza la clasificación y regresión para la predicción de los atributos discretos y continuos. Para los atributos discretos el árbol de decisión realiza una predicción basándose en un conjunto de datos de entrada y utiliza los valores conocidos como estados. Con los atributos de calidad se utiliza la regresión lineal para conocer donde se divide el árbol. [33]

Este se puede dividir de una manera binaria recursiva, en este proceso se logra considerar todas las características del mismo y se prueban en diferentes puntos para realizar una función de costos, con el cual se selecciona la división con el costo más bajo; la función de costo es la que utiliza la clasificación y regresión, pero en ambos casos este encuentra la mayor parte de las ramas homogéneas o las que se componen de grupos con respuestas idénticas, el árbol de decisión solamente se detiene al establecer un número mínimo de

[31] dataprix.com [en línea] análisis discriminante < https://www.dataprix.com/blog-it/mineria-datos/data-mining-analisis-discriminante-caso-sas>

[32] Christopher M. Bishop. Pattern recognition and machine learning: Linear models for regression En: Information Science and Statics. Contents 2. Pages 110.

[33] docs.microsfot.com [en línea] Arboles de decisión <https://docs.microsoft.com/es-es/analysis-services/data-mining/microsoft-decision-trees-algorithm?view=sql-server-2017 >

entradas en el entrenamiento de cada hoja.[34]

2.2.6 Redes neuronales

La red neuronal es la implementación de un aprendizaje automático, en el cual el algoritmo prueba cada posible estado de entrada con cada posible estado del atributo de predicción, con el cual se calcula cada combinación de aprendizaje. “El número de redes incluidas en un modelo de minería de datos depende del número de estados (o valores de atributo) de las columnas de entrada, así como del número de columnas de predicción que usa el modelo de minería de datos y el número de estados de dichas columnas.” [35]

El entrenamiento de las redes neuronales va desde una función x la cual es un vector de variables de entrada y y que representa las variables de salida. Este cuenta con un parámetro para determinar la analogía de la red, se tiene en cuenta el conjunto de entrenamiento de los vectores de entrada $\{xn\}$, donde este va desde $n = 1, \dots, N$, el cual tiene un conjunto de vectores $\{tn\}$ con el cual se minimiza la función de error [36], que se encuentra representada por la Ecuación 6:

Ecuación 6 Redes neuronales

$$E(w) = \frac{1}{2}\sum_{n=1}^{N} ||y(x_{n,}w) - t_n||^2$$

2.2.7 Minería de datos de descriptiva

Esta permite formar grupos de datos con métodos simétricos, los cuales son supervisados o indirectos. Todas las variables que son recolectadas por este método son tratadas al mismo nivel [37]. Además, este se utiliza para extraer datos y lograr proporcionar información de eventos pasados o recientes,

[34] towarddatascience.com [en línea] Arboles de decisión < https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052>

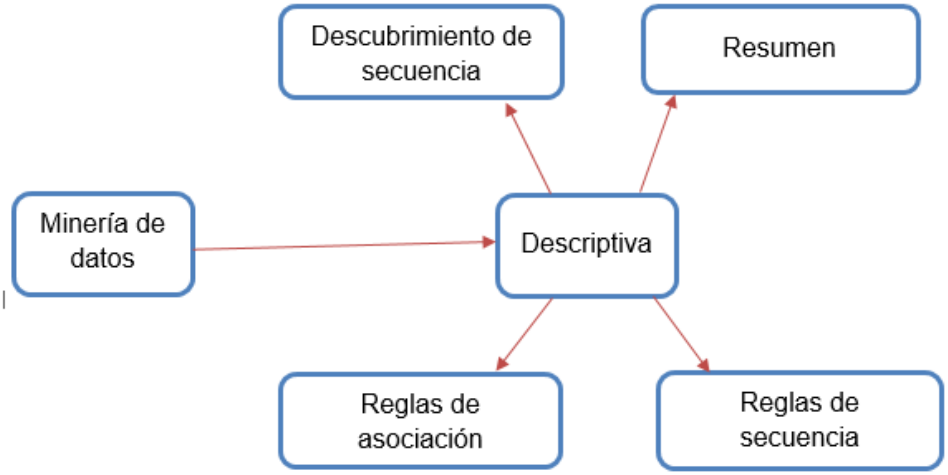
[35] docs.microsfot.com [en línea] Neutral network algorithm <https://docs.microsoft.com/es-es/analysis-services/data-mining/microsoft-neural-network-algorithm?view=sql-server-2017 >

[36] Christopher M. Bishop. Pattern recognition and machine learning: Linear models for regression En: Information Science and Statics. Contents 5. Pages 225 - 236.

[37]Tamps.cinvestav.mx [en línea] Minería de datos descriptiva <https://www.tamps.cinvestav.mx/~hmarin/Mineria/EC2.pdf >

logrando identificar lo que ha sucedido en el pasado con análisis de los datos almacenados realizando una proporción a datos precisos.

La minería de datos descriptiva utiliza correlación, tabulación cruzada, frecuencia, estas técnicas son usadas para llegar a determinar la regularidad de los datos y lograr conocer patrones.[38]
Figura 5 Minería de datos descriptiva



Fuente: Los autores

2.2.8 Clustering

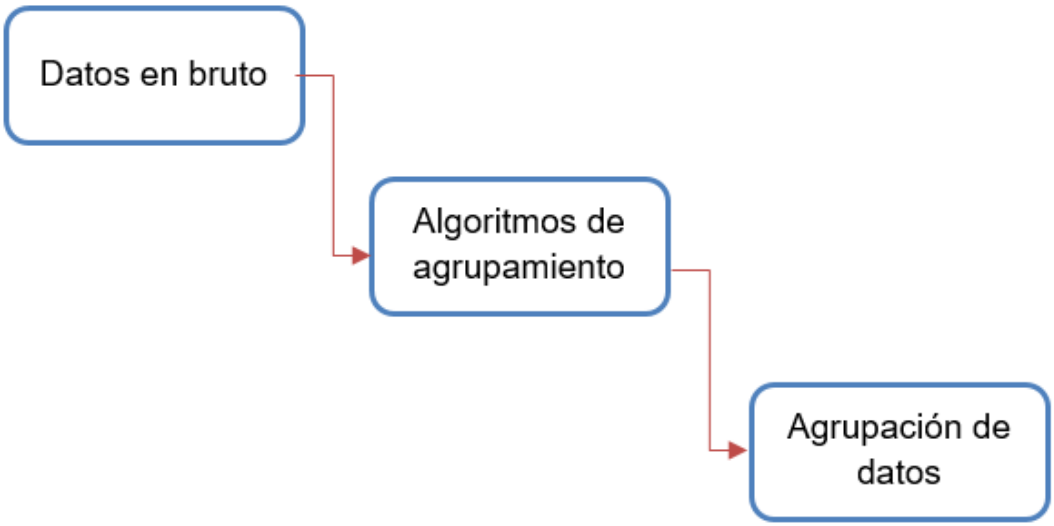
Es una agrupación de una serie de vectores con el cual se manejan los criterios de similitud y distancia. “Identifica de forma automática agrupaciones o clústeres de elementos de acuerdo a una medida de similitud entre ellos”[39], usualmente los vectores del mismo grupo comparten las mismas propiedades.[40] Se puede observar en la Figura 6 las tres etapas básicas de clustering.

[38] Techdifferences.com [en línea] Minería de datos descriptiva< https://techdifferences.com/difference-between-descriptive-and-predictive-data-mining.html >

[39] cs.us.es [en línea] técnicas de clustering < https://www.cs.us.es/~fran/curso_unia/clustering.html>

[40] es.coursea.org [en línea] que es clustering <https://es.coursera.org/lecture/mineria-de-datos-introduccion/que-es-clustering-TMSYv>

Figura 6 Etapas de Clustering



Fuente: Los autores.

La técnica de clustering cuenta con unos objetos específicos en función a las características y similitudes, esta divide los datos de tal manera que un objeto sea estrictamente parte de una partición. Los clústers cuenta con diferentes métodos para su división, uno de estos es basado en particiones en el cual se divide los datos en diferentes subconjuntos indicando que cada grupo puede llegar a tener al menos un objeto y que cada objeto le pertenece a un grupo, el segundo método es basado en la densidad el cual tiene como objetivo la producción de grupos dada en la participación de alta densidad de conjuntos de dato y el tercer método es el jerárquico en el cual se crea una descomposición de los conjuntos dado cada objeto en el dato este cuenta con una clasificación de enfoque de aglomeración en el cual se observa cómo se encuentra ejecutando los objetos o grupos y por último el de enfoque divisivo es en el que no se logra realizar la división o fusión de los objetos con los grupos cuando estos ya se encuentran creados.[41]

2.2.9 Segmentación

La segmentación es la que divide los datos en grupos (Clusters), el cual contiene propiedades similares para su división.[42] Esta ayuda a descubrir las características que se pueden llegar a tener en un conjunto de datos, en este método se detecta y describe las similitudes de tal modo que estas sean

[41] educba.com [en línea] Clustering < https://www.educba.com/what-is-clustering-in-data-mining/ >

[42] docs.microsfot.com [en línea] Algoritmos de minería de datos <https://docs.microsoft.com/es-es/analysis-services/data-mining/data-mining-algorithms-analysis-services-data-mining?view=sql-server-2017>

dirigidas a grupos definiendo los segmentos, llegando así a utilizar información respaldada por los datos, además esta se puede llegar a repetir de tal manera que se adapta a los cambios imperceptibles para lograr notar los cambios que se realicen a lo largo del tiempo [43].

2.2.10 Asociación

Los algoritmos de asociación buscan la diferencia entre los atributos de un conjunto determinado de datos, realizando "correlaciones entre diferentes atributos de un conjunto de datos". [44]

La asociación es una función de minería de datos la que se basa en la probabilidad de la ocurrencia simultánea de unos elementos concurrentes como lo son las reglas de asociación. La regla de asociación se utiliza para analizar las transacciones de ventas, siendo así un comparativo de los clientes a la hora que estos realizan compras, puesto que la asociación se basa en las transacciones realizando la abstracción de un atributo[45].

En la Figura 7 se observa que sucede cuando un comprador realiza la compra de algo, puesto que una de las reglas de asociación es uno de los conceptos importantes del aprendizaje automático que se utiliza en el análisis de la cesta de la compra, ya que en una tienda todas las verduras se encuentran en un mismo pasillo, todos los productos lácteos están juntos y los cosméticos en otro grupo. Este procedimiento hace que el cliente reduzca tiempo en sus compras, pero a su vez le recuerda que artículos relevantes podría estar interesado en comprar, lo que hace que este procedimiento realice ventas cruzadas en el proceso. La asociación consiste en un antecedente y un consecuente, los cuales son una lista de elementos[46].

---

[43] ibm.com [en línea] Segmentation <https://www.ibm.com/support/knowledgecenter/en/SSEPGG_9.7.0/com.ibm.datatools.datamining.doc/miningplan_custseg.html>

[44] docs.microsfot.com [en línea] Algoritmos de minería de datos <https://docs.microsoft.com/es-es/analysis-services/data-mining/data-mining-algorithms-analysis-services-data-mining?view=sql-server-2017>

[45] Docs.oracle.com [en línea] association <https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/market_basket.htm#DMCON009>

[46] Towardsdatascience.com [en línea] Complete guide to association rules <https://towardsdatascience.com/association-rules-2-aa9a77241654>

Figura 7 Asociación

Fuente: Association [en línea]. [Citado el 22 octubre, 2019]. Disponible en internet: < https://blogs.adobe.com/digitalmarketing/wp-content/uploads/2013/08/pic1.jpg >

2.2.11 Análisis exploratorio

El análisis exploratorio en el cual se emplea el análisis de datos para lograr descubrir estructuras subyacentes, entender los conjuntos de datos y detectar valores atípicos o anomalías en los datos. [47]

Este consta de dos maneras generales para su clasificación cruzada, el primer método consta de una manera gráfico o no gráfico y el segundo método es univariado o multivariado. El primer método no gráfico consta del cálculo de probabilidades y en el que el gráfico los representa de manera esquemática, el segundo método univariado solamente se fija en una variable mientras que el multivariados tiene en cuenta dos o más variables[48].

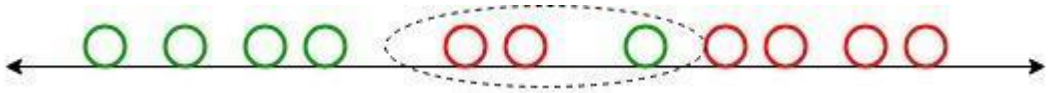
[47] urg.es [en línea] analisis exploratorio < https://www.ugr.es/~batanero/pages/ARTICULOS/anaexplora.pdf >

[48] stat.cmu.edu [en línea] análisis exploratorio < https://www.stat.cmu.edu/~hseltman/309/Book/chapter4.pdf>

2.2.12 Análisis discriminante lineal

El análisis de discriminante lineal (LDA) es una técnica que se utiliza para la reducción de dimensionalidad para los problemas de clasificación, esta es usada para modelar la diferencia entre grupos; ya sea que estos se encuentran separados por dos clases o más. En la *Figura 8* se puede observar que en el momento de usar una sola función se tiene un problema para clasificar, ya que se necesita aumentar el número de características para una buena clasificación [49].

Figura 8 Superposición



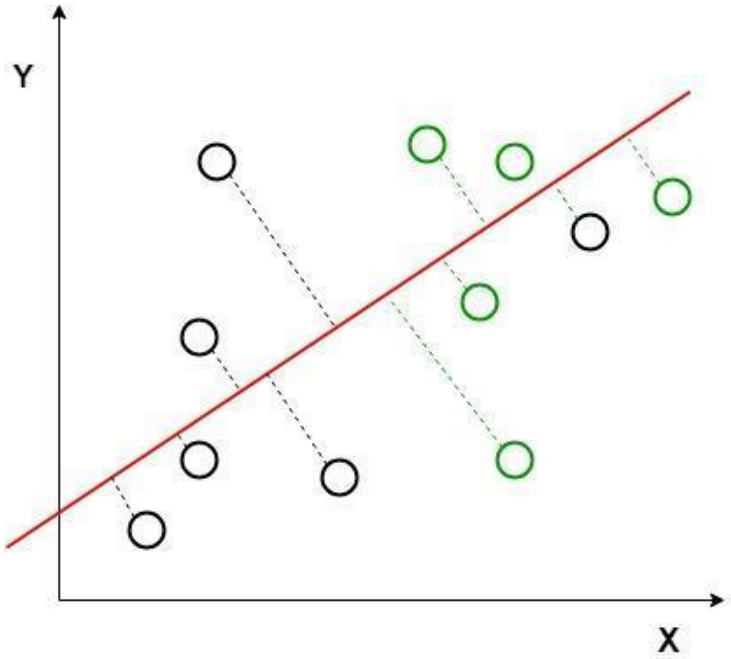
Fuente: Linear discriminant analysis [autor] Raman Disponible en: < https://www.geeksforgeeks.org/ml-linear-discriminant-analysis/>

Como se puede observar en la Figura *8*9 se realiza un aumento de características para realizar un LDA contiene propiedades estadísticas en sus datos para calcular cada clase se utiliza la media y la varianza de la variable, cuando esta se encuentra compuesta por múltiples variables se calcula el gaussiano multivariado [50]. Estimando las propiedades estadísticas a partir de sus datos. Se puede observar en la *Figura 9* que con el aumento de características se crea un nuevo eje con el cual se maximiza la distancia entre las dos clases y se minimiza la variación dentro de la clase.

[49] Geeksforgeeks.org [en línea] análisis discriminante lineal < https://www.geeksforgeeks.org/ml-linear-discriminant-analysis/>

[50] Machinelearningmastery.com [en línea] análisis discriminante lineal <https://machinelearningmastery.com/linear-discriminant-analysis-for-machine-learning/>

Figura 9 LDA



Fuente: Linear discriminant analysis [autor] Raman Disponible en: < https://www.geeksforgeeks.org/ml-linear-discriminant-analysis/>

2.2.13 Análisis discriminante cuadrático

El análisis de discriminante cuadrático (QDA) modela la distribución de los diferentes predictores de x los cuales van por separado en cada una de las clases de respuesta, este utiliza teorema de Bayes para convertir en estimaciones en probabilidad dado el valor x [51]. Este modela y clasifica a y con una combinación no lineal de las variables predictivas de x, QDA se convierte particularmente útil en el momento del conocimiento previo de las clases individuales que se exhiben covarianzas distintas[52].

Ecuación 7 Análisis discriminante cuadrático

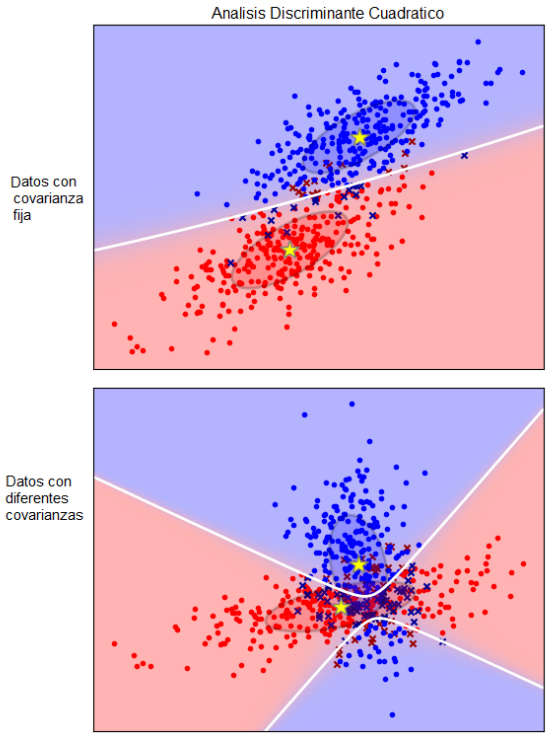
$$\delta_k(x) = -\frac{1}{2}\log\left|\sum k\right| - \frac{1}{2}(x-\mu\, k)^T \sum_k^{-1}(x-\ \mu\, k) + \log\pi\, k$$

[51] Uc-r.github.io [en línea] discriminant analysis <http://uc-r.github.io/discriminant_analysis>

[52]Datascienceblog.net [en linea] linear and quadratic discriminant analysis <https://www.datascienceblog.net/post/machine-learning/linear-discriminant-analysis/>

En la Figura 10 se observa la forma de clasificar los datos por el método de análisis discriminante cuadrático, además que el QDA puede aprender límites cuadráticos, haciéndolo más flexible.

Figura 10 Análisis de discriminante cuadrático



Fuente: Linear and quadratic discriminant analysis Disponible en < https://scikit-learn.org/stable/modules/lda_qda.html>

2.2.14 $\mathbf{F_1 - Score}$

$\mathrm{F_1 - Score}$ es una medida de precisión de una prueba para en puntaje F, en el que se define como la medida armónica ponderada de la precisión y recuperación de la prueba. Esta puntuación se calcula de acuerdo con la ecuacion:

Ecuación 8 $F_1 - Score$

$$F1 = \left(\frac{recall^{-1} + precision^{-1}}{2}\right)^{-1} = 2\frac{precision \times recall}{precision + recall}$$

Es la media armónica entre la precisión y recall, su rango de puntuación es [0 ,1]. Indica cuantas instancias son clasificadas correctamente, así como que tan robusto es, lo que quiere decir que no pierde un número significativo de instancias.

Cuanto mayor es el $\mathrm{F}_1 - \mathrm{Score}$, mejorara el rendimiento del modelo[53], matemáticamente se expresa, $\mathrm{F}_1 - \mathrm{Score}$ intentando encontrar el equilibrio entre precisión y recall.

- **Precisión:** Número de resultados positivos correctos dividido por el número de resultados positivos predichos por el clasificador.
  Ecuación 9 Precisión

$$precision = \frac{verdaderos\ positivos}{verdaderos\ positivos\ +\ falsos\ positivos}$$

- **Recall:** es el número de resultados positivos correctos dividido por el número de todas las muestras relevantes.
  Ecuación 10 Recall

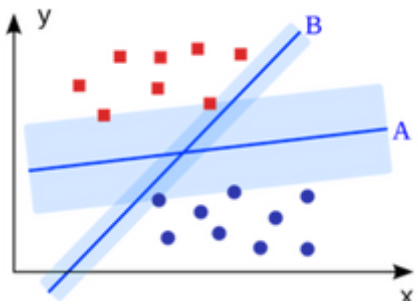
$$recall = \frac{verdaderos\ positivos}{verdaderos\ positivos + falsos\ negativos}$$

2.2.15 Máquina de soporte vectorial

Una máquina de soporte vectorial (SVM – Suppot Vector Machine) es un modelo de aprendizaje automático supervisado[54], en el que se utilizan dos algoritmos de clasificación. Este se define como un clasificador discriminante, el cual se encuentra definido por un hiperplano de separación, en este se encuentra los datos de entrenamiento que es un aprendizaje supervisado. Este algoritmo se genera en un espacio de dos dimensiones o multidimensional, con un hiperplano por el que se encuentra separado en dos partes o más[55].

[53] towardsdatascience.com [en linea]< https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234>
[54] Monkeylearn.com [En línea] <https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/>
[55] Medium.com [En línea] <https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72>

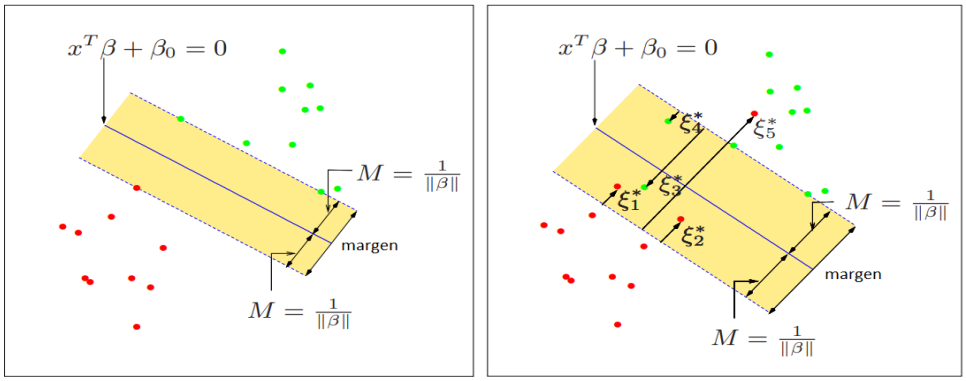
Figura 11 Maquina de soporte vectorial



Fuente: Support Vector Machine [autor] Savan Patel (Theory) Disponible en < https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72>

En la *Figura 12*, en la parte izquierda de la imagen se evidencia la separación de los límites de decisión en la línea continua, mientras que las líneas discontinuas delimitan el sombreado sobre el margen que contiene un máximo de ancho $2\mathrm{M} = \frac{2}{\|\beta\|}$. Mientras que, en la figura de la parte derecha, se logra observar los puntos que no se lograron separar, el cual se considera una superposición; los puntos que se encuentran etiquetados ε_j^* son los que encuentran en el lado equivocado de la margen, por otro lado el margen es maximizado sujeto a un valor total $\sum \varepsilon_i \leq$ contante [56].

Figura 12 Clasificador de máquina de soporte vectorial



Fuente: Trevor Hastie, Robert Tibshirani, Jerome Friedman. The elements of statistical learning, The second edition. En: Chapter 12. Page 418.

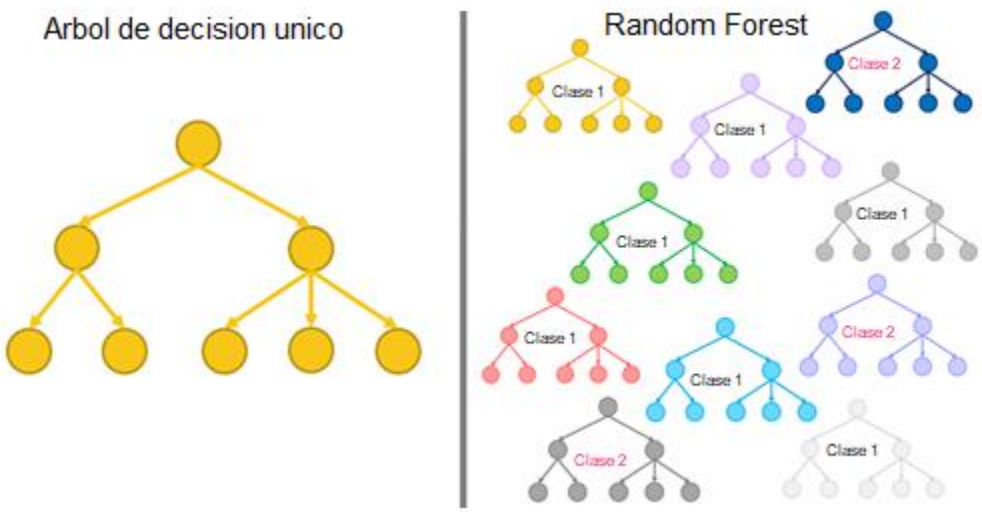
[56] Trevor Hastie, Robert Tibshirani, Jerome Friedman. The elements of statistical learning, The second edition. En: chapter 12. Page 417 – 418.

2.2.16 Random Forest

Random Forest (RF) es un algoritmo de aprendizaje supervisado con el cual se crea y se combina aleatoriamente los múltiples árboles de decisión, además es un meta estimulador que se ajusta a distintos clasificadores y utiliza el promedio para mejorar la precisión de predicción y controlar el sobre ajuste[57].

Las estructuras en los datos crecen lo suficientemente grandes, teniendo en cuenta la baja parcialidad. Los árboles de decisiones son ruidosos, puesto que estos se miden con el promedio, además ya que cada árbol que se genera se distribuye de manera idéntica.

Figura 13 Árbol de decisiones RF



Fuente: From a single decision tree to a random forest [autor] Rosaria Silipo Disponible en: <https://www.dataversity.net/from-a-single-decision-tree-to-a-random-forest/>

La expectativa de un promedio de B, de los arboles es la misma de cualquiera de ellos en la segmentación, esto se da ya que el sesgo de los arboles juntos es igual que la de los arboles individuales y su forma de mejora es a través de la reducción de la varianza. Donde B consta de variables aleatorias y σ^2

[57]scikit-learn.org [en línea]<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>

contiene la varianza de $\frac{1}{B}\sigma^2$. Si las variables son de correlación positiva en pares p, su varianza promedio es la Ecuación 11[58]:

Ecuación 11 Random Forest

$$p\sigma^2 + \frac{1-p}{B}\sigma^2$$

2.2.17 Naive Bayes

Naive Bayes es un modelo probabilístico de aprendizaje automático, el cual se utiliza para la clasificación. Usando el teorema de bayes, se puede encontrar la probabilidad de que ocurra en A, dado que B ya ha ocurrido, donde B es la evidencia y A es la hipótesis[59]. Esta también se encuentra definida por la probabilidad marginal, la cual se encuentra compuesta por variables dadas aleatoriamente de manera bidimensional, la cual se encuentra definida como $P(A) = \sum j\, P(A_i, B_j)$[60]. Dada la probabilidad conjunta y marginal, la probabilidad condicional se define en la Ecuación 12.

Ecuación 12 Probabilidad de Naive Bayes

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Los clasificadores de Naive Bayes se ha convertido en una técnica muy popular, ya que esta es especialmente apropiada cuando la dimensión p del espacio de características es alta, esto hace que la estimación de densidad se convierta en una forma no atractiva. Además, la densidad marginal de la clase individual f_{jk} es estimada por separado utilizando las estimaciones unidimensionales de densidad del núcleo. Si un componente X_j de X es discreto, esto proporciona una manera perfecta de mezclar variables en un vector de características[61].

Ecuación 13 Naive Bayes

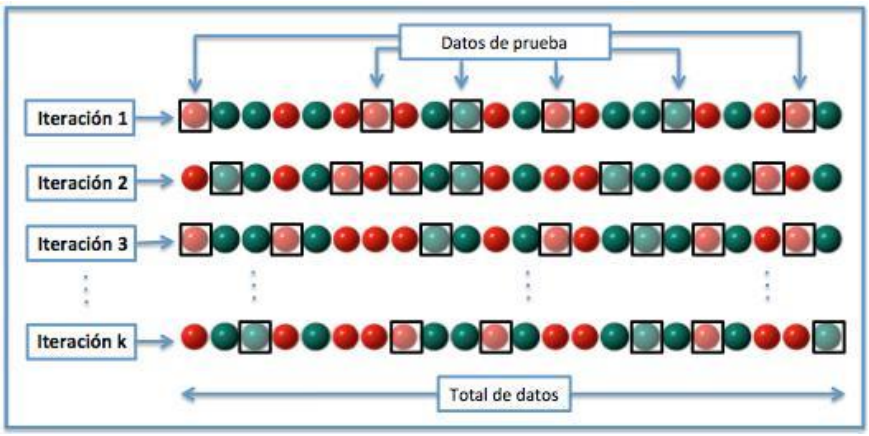
$$f_i(X) = \prod_{k=1}^{p} f_{jk}(X_k)$$

[58] Trevor Hastie, Robert Tibshirani, Jerome Friedman. The elements of statistical learning, the second edition. En: chapter 15. Page 587– 589.
[59] Towards data science [en línea] < https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c>
[60] Naive Bayes [en línea] <https://ccc.inaoep.mx/~emorales/Cursos/NvoAprend/Acetatos/bayes.pdf>
[61] Trevor Hastie, Robert Tibshirani, Jerome Friedman. The elements of statistical learning, the second edition. En: chapter 6. Page 210 - 211.

2.2.18 Validación Cruzada

La validación cruzada, es una técnica utilizada para probar el rendimiento y la efectividad de los modelos de aprendizaje automático. Este también es un procedimiento de muestreó el que se utiliza para evaluar el modelo en donde se obtienen datos limitados, para realizar la validación cruzada como se observa en la Figura 14 se realiza una separación del muestreo y la proporción de los datos en el cual se entrena el modelo. En este enfoque se hace una división aleatoria de los datos en un conjunto de entrenamiento y prueba[62], realizando iteraciones diferentes ya que de esta manera se logra conocer la predicción de un modelo de aprendizaje automático se compara el conjunto de validación y las etiquetas reales de los puntos de datos. Además, para poder realizar la reducción de la varianza, se debe realizar distintas iteraciones de validación cruzada utilizando distintos conjuntos de entrenamiento.

Figura 14 Validación cruzada



Fuente: Cross Validation [autor] Juan Gabriel Golima Disponible en: <http://rstudio-pubs-static.s3.amazonaws.com/423338_5b4dc6a938144a3b8ab2ce01fe8be14f.html>

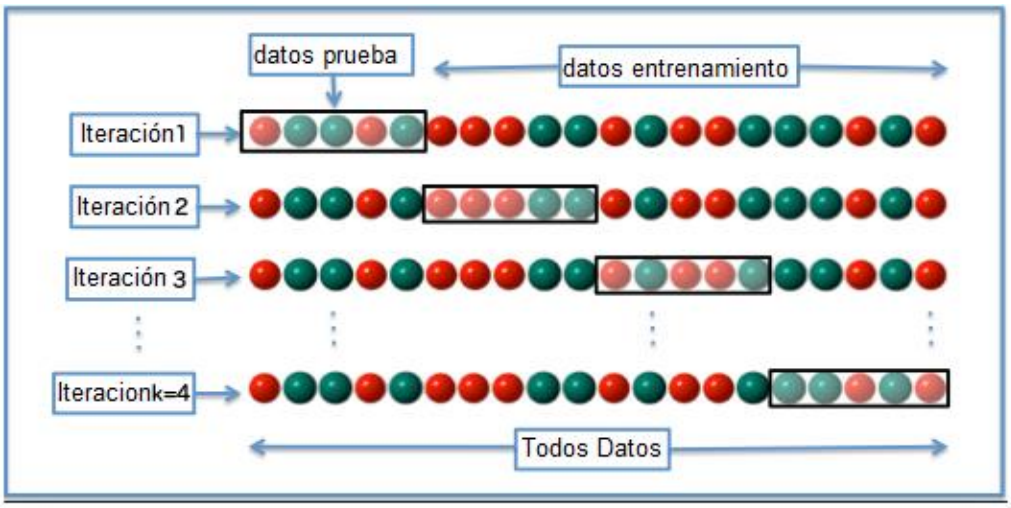
2.2.19 K-Folds

K-Folds es un modelo que es menos sesgado realizando su comparación con otros métodos, como se observa en la Figura 15, este garantiza la obtención de un conjunto de datos original con el que se tiene la posibilidad de obtener un conjunto de entrenamiento y otro de prueba, este es un buen enfoque ya que se tiene datos de entrada limitados. Su primer paso es dividir todos los

[62] Towards data science [en línea] < https://towardsdatascience.com/why-and-how-to-cross-validate-a-model-d6424b45261f>

datos al azar en K pliegues, donde el valor más alto de K se da en un modelo sesgado y donde el valor más bajo de K es similar al enfoque dividido de prueba, segundo se realiza los ajustes de $K-1$ y se valida el modelo, repitiendo el procedo de K-Folds[63].

Figura 15 K Folds



Fuente: K-Folds Cross Validation [autor] Sanjay.M Disponible en: <https://towardsdatascience.com/why-and-how-to-cross-validate-a-model-d6424b45261f >

2.2.20 TF-IDF

El procesamiento de lenguaje natural, es un subcampo de inteligencia artificial el cual se ocupa de la compresión y el procesamiento del lenguaje humano, con TF-IDF se le asigna la importancia que se obtendrá de cada palabra.

TF es el término de frecuencia, con el cual se mide el número de veces que llega a repetirse una palabra en un documento, dividido por el número total de palabras en el documento[64], donde $\mathrm{tf}_{i,j}$ es el número de ocurrencias de i en j y $\mathrm{n}_{i,j}$ es el termino de frecuencia de i en el documento de j.

Ecuación 14 Termino de Frecuencia

$$tf_{i,j} = \frac{n_{i,j}}{\sum k\, n_{i,j}}$$

IDF es la frecuencia de datos inversa la cual determina el peso de las palabras en todo el documento, este es el registro de la cantidad de documentos dividido por la cantidad de documentos que contiene w. N es el número total de documentos y df_t es el número de documentos que contiene el término.

[63] Towardsdatascience.com [en línea] <https://towardsdatascience.com/natural-language-processing-feature-engineering-using-tf-idf-e8b9d00e7e76>

[64] Towardsdatascience.com [en línea] <https://towardsdatascience.com/natural-language-processing-feature-engineering-using-tf-idf-e8b9d00e7e76>

Ecuación 15 Frecuencia de datos inversa

$$idf(w) = log(\frac{N}{df_t})$$

Por último, la ecuación de TF-IDF es la multiplicación de estas dos, siendo $w_{i,j}$ el peso por ficha de i en el documento j, $\mathrm{tf_{i,j}}$ es el número de ocurrencias de i en j:

Ecuación 16 TF-IDF

$$w_{i,j} = tf_{i,j} \,.\, log\, \frac{N}{df_i}$$

2.2.21 N-Gram

N-Gram es un modelo estadístico, con el que se asigna las probabilidades de la secuencia de las palabras. Este modelo tiene como la secuencia de N palabras, según esa noción, un unigrama es una secuencia de una palabra, un 2-gram mejor conocido como un bigrama es la ejecución de dos palabras, un ejemplo de esto es “gire su” o “su tarea”.

El modelo de bigrama es la aproximación de probabilidad de una palabra dada todas las palabras anteriores, al usar sola la probabilidad condicional de una palabra anterior[65]. el modelo bigrama, por ejemplo, aproxima la probabilidad de una palabra dadas todas las palabras anteriores $P(w_n|w_1^{n-1})$ al usar solo la probabilidad condicional de la palabra anterior $P(w_n|w_{n-1})$.

En otras palabras, en lugar de calcular la probabilidad aproximada $P(La|\mathrm{Que})$, cuando se usa un modelo bigrama para predecir la probabilidad condicional de la siguiente palabra, se realiza la siguiente ecuación[66]:

Ecuación 17 Bigrama

$$P(w_n|w_1^{n-1}) = P(w_n|w_{n-1})$$

[65]Towardsdatascience.com [en línea] < https://towardsdatascience.com/introduction-to-language-models-n-gram-e323081503d9>

[66] N-Gram [en línea] <https://www.slideshare.net/shkulathilake/nlpkashkngrams>

2.2.22 Knime

Knime es una plataforma de minería de datos la cual permite el desarrollo en un software de entorno visual. Knime es una innovación basada en datos, diseñada para descubrir el potencial oculto en los datos, para lograr predecir nuevos futuros y realizar de manera visual los procedimientos de análisis[67]. Figura 16.

Figura 16 Knime

Fuente: Knime Disponible en < https://www.knime.com/>

2.2.23 Kernel

Se puede considerar que los métodos de kernel proporcionan explícitamente estimaciones de la función de regresión o expectativa condicional al especificar la naturaleza de su local, y de la clase de funciones regulares ajustadas localmente. Los vecinos locales se especifican mediante una función de núcleo $k_{\lambda}(x_{0,}x)$, ponderado a los puntos x en una región de x_0, en el núcleo gaussiano el cual tiene un peso de función basada en la densidad de la función. Este asigna pesos que mueren exponencialmente con su cuadrado de distancia ecluideana de x_0. El parámetro λ corresponde a la varianza de la densidad gaussiana y controla el ando de los vecinos. La representación se puede observar en la Ecuación 18 *Kernel*.

Ecuación 18 Kernel

$$k_{\lambda}(x_{0,}x) = \frac{1}{\lambda} exp\left[-\frac{\|x - x_0\|^2}{2\lambda}\right]$$

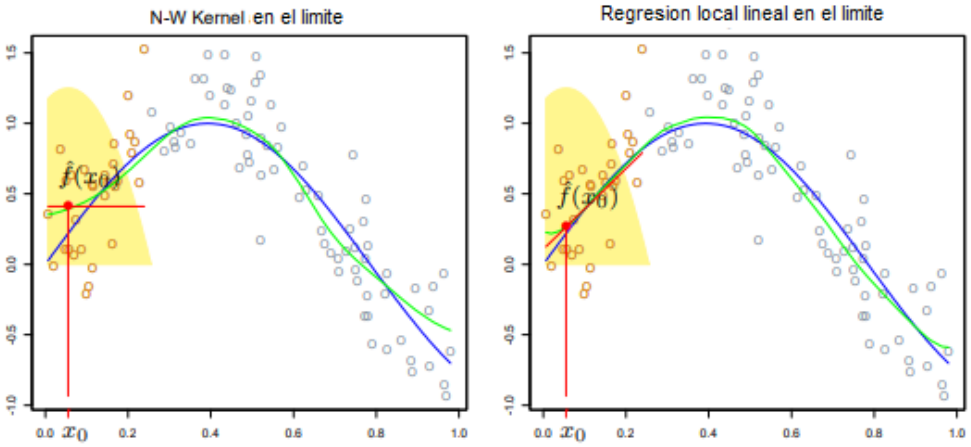
Kernel calcula un producto interno en un espacio de características de alta dimensión y se usa para el modelado no lineal regularizado[68].

[67] Knime.com [en línea] <https://www.knime.com/about>

[68] Trevor Hastie, Robert Tibshirani, Jerome Friedman. The elements of statistical learning, the second edition. En: chapter 6. Page 191 – 200.

Kernel Lineal: En kernel lineal se progresa con la medida móvil bruta a una variación suave a un promedio ponderado mediante el uso local de la ponderación del núcleo. Los promedios ponderados localmente pueden estar muy sesgados en los límites del dominio.

El promedio ponderado localmente tiene problemas de sesgo en o cerca de los límites del dominio. La verdadera función es aproximadamente lineal, pero la mayoría de las observaciones de los vecinos tienen una medida más alta que el punto objetivo, así que, a pesar de la ponderación, su media estará sesgada hacia arriba. Al instalar un local en regresión lineal ponderada, este sesgo se elimina a primer orden.
Figura 17 Kernel lineal



Fuente: Trevor Hastie, Robert Tibshirani, Jerome Friedman. The elements of statistical learning, the second edition. En: chapter 6. Page 195

Debido a la asimetría en el núcleo de la región, en el momento que las líneas tienen una posición derecha, en lugar de constantes localmente, se puede eliminar el sesgo exactamente al primer orden. En realidad, este sesgo puede estar presente en el interior del dominio también, si los valores de x no están igualmente espaciados. La regresión ponderada localmente resuelve un problema separado de mínimos cuadrados ponderados en cada punto objetivo x_0[69], donde $y_i - \propto (x_0) - \beta(x_0)x_i$ se conoce como la función de activación, puesto que $\propto$ es la direccion y los terminos de sesgo son β, el cual tiene que ser determinado por su estimación, y $\searrow$ corresponde a la varianza de la densidad gaussiana.

[69] Trevor Hastie, Robert Tibshirani, Jerome Friedman. The elements of statistical learning, the second edition. En: chapter 6. Page 191 – 200.

Ecuación 19 Kernel Lineal

$$\sum_{i=1}^{N} k \succ (x_0, x_i)[y_i - \propto (x_0) - \beta(x_0)x_i]^2$$

Kernel RBF:

El método gaussiano RBF el cual traduce función de base radial, es una función cuyo valor depende de la distancia desde el origen o desde algún punto. Usando la distancia en el espacio original se calcula el producto de puntos, su similitud de X_1 y X_2. Donde C es el inverso de la fuerza de la regularización, ya que a medida que aumenta el valor de C, el modelo obtiene sobreajustes y a medida que el valor de C disminuye, el modelo se adapta. Da el valor de comportamiento a medida que aumenta el valor de esta misma el modelo obtiene sobre ajustes[70].

Ecuación 20 Kernel RBF

$$C(X_1, X_2) = e^{(-\gamma||X_1 - X_2||^2)}$$

Kernel Polinomial:

Con la solución $\hat{f}(x_0) = \hat{\propto}(x_0) + \sum_{j=1}^{d} \hat{\beta}(x_0)x_0^j$. Donde esta es una expansión del sesgo que solo tendrá componentes de grado $d + 1$ y más alto, los ajustes lineales tienden a estar sesgados en regiones de curvatura de la función verdadera, un fenómeno conocido como recortar las colinas y llenar los valles.

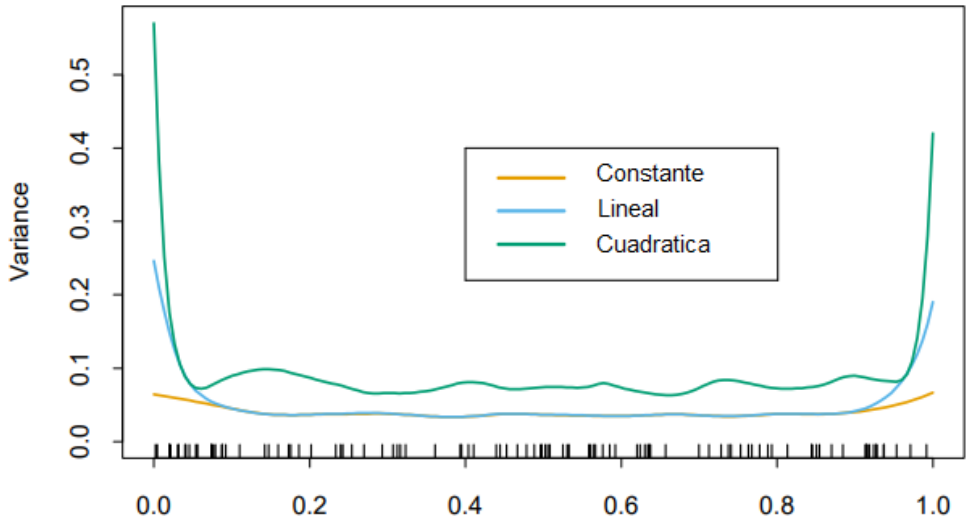
Ecuación 21 Kernel Polinomial

$$\min_{\propto(x_0), \beta_j(x_0),\ j=1,\dots,d} \sum_{i=1}^{N} K_{\succ}(x_0, x_i)[y_i - \alpha(x_0) - \sum_{j=1}^{d} \beta_j(x_0)x_i^j]^2$$

La regresión cuadrática generalmente es capaz de corregir este sesgo, que cuenta con una penalización por esta reducción de sesgo, y esa varianza es el mayor ajuste. Suponiendo que el modelo $y_i = f(x_i) + \varepsilon_i$ con ε_i independiente e idénticamente distribuido con media cero y varianza σ^2 donde $Var\left(\hat{f}(x_0)\right) = \sigma^2 \|l(x_0)\|^2$, donde $\|l(x_0)\|$ es el vector del núcleo equivalente a un peso de x_0.

[70] SVM with kernel [en línea] <https://towardsdatascience.com/support-vector-machines-svm-c9ef22815589>

Figura 18 Kernel Polinomial



Fuente: Trevor Hastie, Robert Tibshirani, Jerome Friedman. The elements of statistical learning, the second edition. En: chapter 6. Page 198.

La varianza de la función $\|l(x)\|^2$ para una constante local lineal y regresión cuadrática, para un núcleo de tres cubos de ancho de banda métrico $(\lambda = 0.2)$. Los ajustes lineales locales pueden ayudar a sesgo dramáticamente en los límites en una penalización en la varianza. Los ajustes cuadráticos locales hacen poco en los límites del sesgo, pero aumentan la varianza.
Los ajustes cuadráticos locales tienden a ser más útiles para reducir el sesgo debido a la curvatura en el interior del dominio.
El análisis asintótico sugiere que los polinomios locales de grado impar dominan los de grado uniforme. Esto se debe en gran parte al hecho de que asintóticamente, está dominado por los efectos de limite[71].

2.2.24 Matriz de confusión

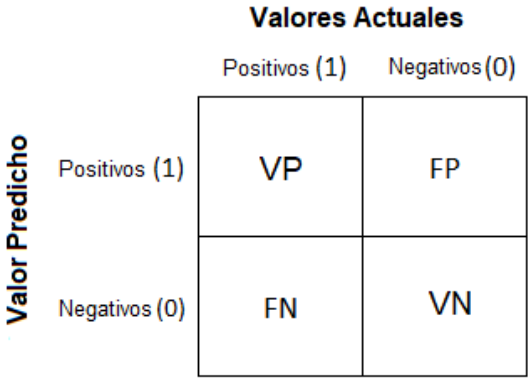
Es una medida de rendimiento para el problema de clasificación de aprendizaje automático, donde la salida puede ser de dos o más clases. Además es una herramienta la cual permite la visualización del desempeño del algoritmo que se emplea en aprendizaje supervisado, cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila es la instancia en la clase real[72].La precisión de la matriz se puede calcular tomando como promedio los valores que se encuentran en la diagonal principal, es decir en la

[71] Trevor Hastie, Robert Tibshirani, Jerome Friedman. The elements of statistical learning, the second edition. En: chapter 6. Page 191 – 200.

[72] Matriz de confusion [en línea] <https://www.juanbarrios.com/matriz-de-confusion-y-sus-metricas/>

precisión[73].

Figura 19 Matriz de confusión



Fuente: Confusion matrix [autor] Sarang Narkhede Disponible en: <https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62>

La matriz de confusión multiclase diferencia del proceso de los problemas de clasificación binaria, no tiene que elegir un umbral de puntuación para realizar predicciones. La respuesta predicha es la clase con la puntuación máxima predicha. Las métricas que se utilizan en la multiclase son las mismas que se utilizan en el caso de clasificación binaria. La métrica se calcula para cada clase al procesarla como un problema de clasificación binaria después de agrupar todas las otras clases como pertenecientes a la segunda clase[74].

2.2.25 Exactitud

La exactitud es una métrica la cual evalúa modelos de clasificación, esta es el número de predicciones correctas y el número total de muestras de entrada[75].

Ecuación 22 Exactitud Probabilidad

$$Exactitud = \frac{Numero\ de\ predicciones\ correctas}{Numero\ total\ de\ predicciones}$$

Tiene como buen funcionamiento el mismo número existente de muestras que pertenecen a cada clase. Además de esto esta tiene una clasificación binaria, la cual se puede calcular en términos de positivos y negativos, donde vp son los verdaderos positivos, vn los verdaderos negativos, fp los falsos positivos y fn los falsos negativos[76].

[73] Confusion matrix [en línea] <https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62>
[74] Docs.aws.amazon.com [en línea] <https://docs.aws.amazon.com/es_es/machine-learning/latest/dg/multiclass-classification.html>
[75] Metrics to evaluate your machine learning [en línea] <https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234>
[76] Machine learning [en línea] <https://developers.google.com/machine-learning/crash-course/classification/accuracy>

Ecuación 23 Exactitud

$$Exactitud = \frac{vp + vn}{vp + vn + fp + fn}$$

En la clasificación de múltiples etiquetas, la función devuelve la precisión del subconjunto. Si el conjunto completo de etiquetas pronosticadas para una muestra coincide estrictamente con el conjunto real de etiquetas, la precisión del subconjunto tendría un valor de 1, de lo contrario su valor es 0.
Si $\hat{y}_i$es el valor predicho de $i-th$ muestra y y_i es el valor verdadero correspondiente, entonces la fraccion de predicciones correctas sobre $n_{samples}$ [77]se define en la Ecuación 24:
Ecuación 24 Multiclase

$$Exactitud(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} 1(\widehat{y_i} = y_i)$$

2.2.26 Macro promedio

Un macro promedio calcula la métrica independiente en cada clase, de esta manera se puede calcular el conjunto de todas las etiquetas. Siendo esta una medida de evaluación binaria, la cual se calcula basada en el número de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos. Este se da en el conjunto de etiquetas, en la que se utiliza el conjunto de datos de entrenamiento. Además, el conjunto de este puede diferir de etiquetas en los datos de la prueba, a su vez el macro promedio otorga el mismo peso a cada clase[78].

2.2.27 Hiperparámetro

Los hiperparámetros son los valores de configuraciones que son utilizadas en el proceso de entrenamiento, estos valores generalmente no se obtienen de los datos, por lo que suelen ser indicados por el científico de datos. Al entrenar un modelo de aprendizaje automático se fijan los valores de los hiperparámetros para que con esto se obtengan los parámetros, algunos de estos pueden ser: El radio de aprendizaje en el algoritmo del descenso del gradiente, el número de vecinos K-vecinos más cercanos, la profundidad máxima en un árbol de decisión[79].

[77] Scikit-learn.org [en línea] <https://scikit-learn.org/stable/modules/model_evaluation.html#accuracy-score>
[78] Semanticscholar [en línea] <https://pdfs.semanticscholar.org/1d10/6a2730801b6210a67f7622e4d192bb309303.pdf>
[79] Analytics Lane [en línea] <https://www.analyticslane.com/2019/12/16/cual-es-la-diferencia-entre-parametro-e-hiperparametro/>

2.2.28 Promedio de ponderación

El promedio ponderado es un cálculo, que se da cuando se multiplica el peso asociado con un evento o resultado particular. También tiene en cuenta los diversos grados de importancia de los números en un conjunto de datos, al calcular un promedio ponderado, cada número en el conjunto de datos se multiplica por un peso predeterminado, ya que este determina la importancia relativa de cada uno de los puntos de los datos[80].

[80] Investopedia [en línea] https://www.investopedia.com/terms/w/weightedaverage.asp

2.3 MARCO CONCEPTUAL

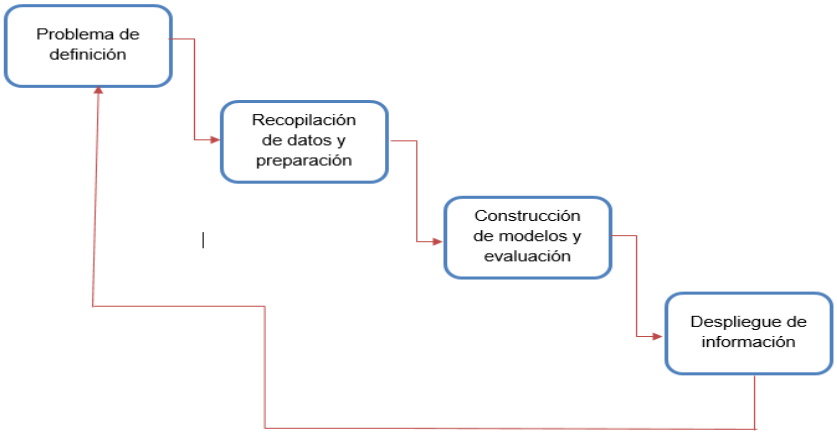
En esta sección se describen los conceptos referentes a la investigación

2.3.1 Minería de datos

Este es el proceso de detectar información, el cual se procesa de grandes conjuntos de datos. La minería de datos utiliza el análisis matemático el cual logra segmentar los datos, para lograr conocer las tendencias y evaluar la probabilidad de eventos futuros [81].

La minería de datos utiliza algoritmos matemáticos para segmentar los datos, esta también se conoce como descubrimiento del conocimiento en datos, manejando un patrón automático para la creación de información procesable. Este se logra mediante la construcción de modelos algorítmicos para actuar sobre un conjunto determinado de datos[82]. Se puede observar en la Figura 20 el proceso que tiene la minería de datos en sus diferentes fases, en su fase de recopilación de datos y preparación se realiza el acceso a los datos del muestreo y la transformación de estos, en su fase de construcción de modelos y evaluación, este crea el modelo lo evalúa y a su vez lo prueba, en su última fase de despliegue de conocimiento aplica el modelo con sus reportes personalizados.

Figura 20 Proceso de minería de datos



Fuente: The data mining process Disponible en: <https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm#DMCON046 >

[81] docs.microsfot.com [en línea] Conceptos de minería de datos <https://docs.microsoft.com/es-es/analysis-services/data-mining/data-mining-concepts?view=sql-server-2017>

[82] docs.oracle.com [en línea] Conceptos de minería de datos < https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm#DMCON046 >

2.3.2 Minería de texto

Es el análisis cualitativo en el cual se realiza una extracción de información del texto, para identificar de manera clara las ideas o conceptos claves que contiene el texto en el cual se pueden agrupar en una serie de categorías apropiadas. Este análisis se puede realizar en un texto de cualquier longitud y tipo. [83]

La minería de texto es la agrupación de texto o palabras que se han extraído de un documento, esta es la unidad de texto que se puede manipular y analizar, con la minería de texto se puede realizar la extracción de términos, búsqueda de palabras y temas[84]; estos pueden venir en datos mixtos en cual incluye contenido estructurado y no estructurado. Este se les aplica a los algoritmos de Bayes, modelos lineales generalizados, Máquinas de Soporte Vectorial, A priori, entre otros. Con estos se realiza la clasificación, el agrupamiento y la extracción de características [85].

2.3.3 Análisis de sentimientos

El procesamiento del lenguaje natural, es el que realiza el seguimiento del estado de ánimo de los usuarios frente a un tema o un producto en particular. El análisis de sentimientos es la construcción de un modelo para recoger y categorizar las diferentes opiniones [86]. Además, es la extracción de texto el cual identifica información subjetiva de ciertas categorías, para predecir si el sentimiento subyacente es positivo, negativo o neutral, combinando el proceso del lenguaje natural y las técnicas de aprendizaje automático[87].

El análisis de sentimientos realiza la división de cada documento de texto en componentes, como los son: las oraciones, frases y tokens la cual es una instancia de un tipo de texto dado. Identificando cada frase que compone el sentimiento y asignando una puntuación a cada frase [88].

---

[83] ibm.com [en línea] Conceptos de minería de texto < https://www.ibm.com/support/knowledgecenter/es/SS3RA7_18.1.1/ta_guide_ddita/textmining/shared_entities/tm_intro_tm_defined.html >

[84] Docs.oracle.com [en línea] About text mining < https://docs.oracle.com/database/121/DMPRG/GUID-3E60BDD1-DE22-494F-8B6D-C73A03EDD01B.htm#DMPRG778 >

[85] Docs.oracle.com [en línea] Text mining <https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/text.htm#BCEDHEDD >

[86] arimetrics.com [en línea] Análisis de sentimientos <https://www.arimetrics.com/glosario-digital/analisis-de-sentimiento>

[87] towardsdatascience.com [en línea] Sentiment analysis: concept, analysis and applications < https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17>

[88] lexalytics.com [en línea] Sentiment analysis explained <https://www.lexalytics.com/technology/sentiment-analysis>

2.3.4 Conjunto de datos

Este es el histórico de los datos que se usa para realizar el entrenamiento del sistema con el cual se logra detectar los diferentes patrones que se componen de las instancias, características y propiedades [89]. El conjunto de datos hace referencia a un archivo que contiene uno o más registros, ya sea desde registros médicos o un registro de seguros. Este cuenta con tres tipos de conjunto de datos, el primero es el secuencial el cual realiza los registros de los elementos de los datos que se almacenan consecutivamente estos pueden ser la lista alfabética de nombres, la segunda es el particionamiento que consta de un directorio que contiene la dirección de cada miembro del sistema operativo y por último es el conjunto de datos VSAM secuenciado de claves del método de acceso que contiene datos de almacenamiento virtual [90].

2.3.5 Modelos de predicción

Los modelos de predicción contienen propiedades los cuales definen el modelamiento y sus metadatos, en los cuales se analizan el nombre, la descripción, la fecha de procesamiento, los permisos y los filtros que se utilizan para el tratamiento [91].

Son los que afinan el análisis de los datos, ya que cuanto más se entrena un modelo, más preciso se vuelve el análisis de riesgos. Para el entrenamiento de los modelos predictivos se usa de dos maneras, la primera es el modelo de detección de anomalías en el cual se realiza su entrenamiento automático cuando se alimenta los datos de acceso histórico y el segundo es el modelo de clasificación de fraude este se entrena sobre el hallazgo de los investigadores de un fraude [92]. En la Figura 211 se observa el procesamiento que se tiene en los modelos de predicción utilizando un algoritmo.

---

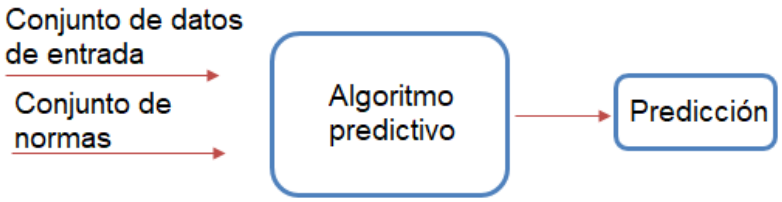
[89] cleverdata.io [en línea] conjunto de datos <https://cleverdata.io/conceptos-basicos-machine-learning/>

[90] Ibm.com [en línea] what is a data set? < https://www.ibm.com/support/knowledgecenter/zosbasics/com.ibm.zos.zconcepts/zconc_datasetintro.htm>

[91] docs.microsfot.com [en línea] Modelo de minería de datos <https://docs.microsoft.com/es-es/analysis-services/data-mining/mining-models-analysis-services-data-mining?view=sql-server-2017#bkmk_mdlDefine>

[92] docs.oracle.com [en línea] Predictive analysis <https://docs.oracle.com/cd/E28280_01/admin.1111/e14568/predict.htm#AAMAD5159>

Figura 21 Modelo predictivo



Fuente: Prediction models: traditional versus machine learning [autor] Jitendra Subramanyam Disponible en <https://blogs.gartner.com/jitendra-subramanyam/prediction-models-traditional-versus-machine-learning/>

2.3.6 Modelo supervisado

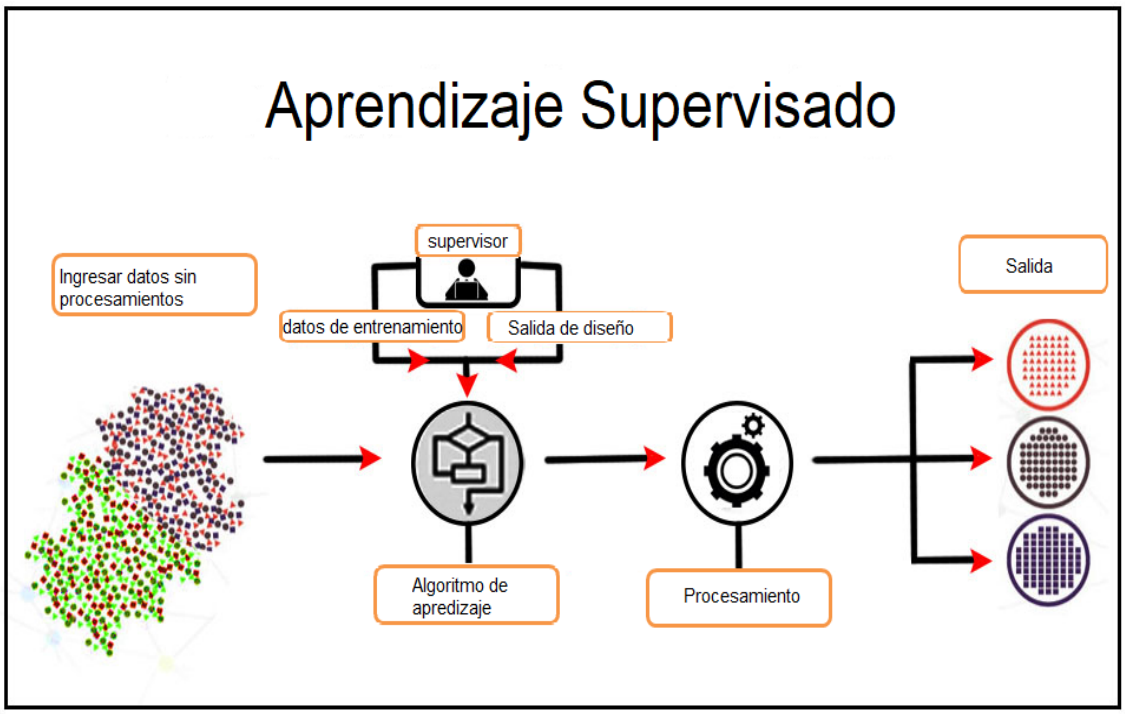
El aprendizaje supervisado consiste en realizar predicciones a futuro que se basan en comportamientos o características que se han visto en los datos que se han almacenado. Este nos permite buscar patrones en los datos históricos relacionados [93].

El modelo supervisado se basa en datos históricos, este tiene como objetivo realizar predicciones con precisión sobre la clase objetivo para cada registro de datos nuevos, es decir datos que no se encuentran en los históricos. Para la clasificación de los datos se tiene en cuenta los valores objetivos, teniendo en cuenta las diferentes técnicas para encontrar relación entre los atributos del predictor y el valor de los datos compilados, su clasificación puede tener dos objetivos uno binario en el cual solo se tienen aquellos que toman dos valores y los multiclase que son los que toman más de dos valores. Este modelo cuenta con tres algoritmos de clasificación, el primero es el algoritmo del árbol de decisión el cual brinda transparencia al modelo y proporciona velocidad y escalabilidad, la segunda es el algoritmo de Naive Bayes el cual se utiliza para los problemas de clasificación binaria y multiclase realizando una escala lineal en el número de predictores y filas y por último el algoritmo de máquina de soporte vectorial en el cual se realiza clasificación y regresión, en el cual se proyectan datos de entrada construyendo un modelo lineal, separando las

[93] cleardata.io [en línea] Conceptos básicos de Machine Learning <https://cleverdata.io/conceptos-basicos-machine-learning/>

clases objetivo con el margen más amplio posible [94]. En la Figura 22 se puede observar procesamiento supervisado que tiene este modelo.

Figura 22 Aprendizaje supervisado



Fuente: Supervised learning model Disponible en: <https://www.datavedas.com/supervised-models/>

2.3.7 Jurisdicción especial para la paz (JEP)

La JEP fue creada por el acuerdo de paz entre el Gobierno Nacional y las Farc-Ep, esta tiene como función administrar la justicia transicional y lograr conocer los delitos que se cometieron en el marco del conflicto armado. La cual fue creada para lograr satisfacer los derechos de las víctimas en la justicia, en la que se les ofrece verdad y contribuir con su reparación [95].

[94] Docs.oracle.com [en línea] supervised data mining < https://docs.oracle.com/cd/B19306_01/datamine.102/b14339/3predictive.htm#i1005885>

[95] jep.gov.co [en línea] Jurisdicción especial para la paz <https://www.jep.gov.co/Paginas/JEP/Jurisdiccion-Especial-para-la-Paz.aspx >

2.4 ESTADO DEL ARTE

El proceso que se realizó para el estado del arte fue revisar diferentes documentos publicados en internet entre los años de 2010 y 2018, se descubrió que, para la extracción y clasificación de tweets, se han realizado mediante diferentes tipos de métodos y distintos procesos.

En el artículo de Ema Kusena, Mark Strembeck, para poder realizar el análisis de sentimientos respecto a las elecciones presidenciales austriacas se captaron alrededor de 343.645 tweets, realizando la clasificación de estos en negativos y positivos, esto permitió una identificación de la polarización para llegar a conocer el ganador a la candidatura [96].

En la investigación de Efthymios Kouloumpis, Theresa Wilson, Johanna Moore, se realizó un modelo supervisado para etiquetar diferentes cantidades de tweets de cinco diferentes conjuntos de datos, con el uso de hashtags para recopilar datos y realizar el entrenamiento y clasificación de los sentimientos entre positivos y negativos [97].

En la investigación de Eric S.Tellez, Sabino Miranda-Jiménez, Mario Graff, Daniela Moctezuma, Oscar S.Siordia, Elio A.Villaseñor, identificaron un gran conjunto de combinaciones que realizan lematización, derivación y eliminación de identidades. Realizando esquemas de ponderación de tokens que tienen un mayor impacto que en la precisión de un clasificador de (Support Vector Machine) entrenándolo en dos conjuntos de datos de palabras en español, utilizando todas las combinaciones de transformaciones de texto y sus respectivos parámetros[98].

En el artículo de Marcela Mayumi Mauricio Yagui, Luís Fernando Monsores Passos Maia, Jonice Oliveira, Adriana S. Vivacqua, se realizó el análisis del comportamiento de los usuarios en la red social Twitter, para llegar a conocer la reacción de la gente frente a un artículo en particular. Se tuvo en cuenta la interrelación y el PageRank para lograr identificar que usuario tuvo influencia sobre los demás, además se analizaron que opiniones y sentimientos fueron

[96] Ema Kušena, Mark Strembeck [año de publicacion] 2018 Politics, sentiments, and misinformation: An analysis of the Twitter discussion on the 2016 Austrian Presidential Elections. Disponible en < https://www-sciencedirect-com.ucatolica.basesdedatosezproxy.com/science/article/pii/S2468696417301088>

[97] Efthymios Kouloumpis, Theresa Wilson, Johanna Moore [año de publicacion] 2010 Twitter Sentiment Analysis:The Good the Bad and the OMG!. Disponible en < https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2857/3251>

[98] Eric S.Tellez, Sabino Miranda-Jiménez, Mario Graff, Daniela Moctezuma, Oscar S.Siordia, Elio A.Villaseñor [año de publicacion] 2017 A case study of Spanish text transformations for twitter sentiment analysis. Disponible en < https://www-sciencedirect-com.ucatolica.basesdedatosezproxy.com/science/article/pii/S0957417417302312>

más frecuentes, etiquetaron1500 tweets [99].

En esta investigación de Ankita, Nabizath Saleenaa, se clasificaron los tweets entre positivos y negativos utilizando Naive Bayes, Random Forest, SVM y regresión. Mostrando la mejora que se generó en el sistema con el desempeño en la clasificación de sus conjuntos [100].

En el artículo de Samah Mansour, realizaron el análisis mediante la red social Twitter con minería de texto y el análisis de sentimientos para lograr conocer si existe una diferencia entre las personas de los países occidentales y los orientales, buscando una comparación entre estos dos y así mismos clasificándolos de acuerdo con la diferencia que existe entre los dos países[101].

En la investigación de Carlos Arcila-Calderón, Félix Ortega-Mohedano, Javier Jiménez-Amores y Sofía Trullenque, realizaron análisis supervisado de sentimientos políticos en español los cuales son clasificados en tiempo real basado en aprendizaje automático, dividiendolo en big data, exalogia, exadata y exalytics [102].

Por último, Lina Andrea Torres Samboni realizo análisis de sentimientos sobre el posconflicto colombiano utilizando herramientas de minería de texto, utilizando business understanding, data understanding, modeling, evaluation y deployment, clasificando los sentimientos se etiquetaron entre negativo y positivo, finalmente los categorizo entre guerrilla, rssMedia, socialmedia, politicans [103].

---

[99] Marcela Mayumi Mauricio Yagui, Luís Fernando Monsores Passos Maia, Jonice Oliveira, Adriana S. Vivacqua [año de publicacion] 2018 Data mining of social manifestations in Twitter: Analysis and aspects of the social movement "Bela, recatada e do lar" (Beautiful, demure and housewife) Disponible en < http://web.a.ebscohost.com.ucatolica.basesdedatosezproxy.com/ehost/pdfviewer/pdfviewer?vid=1&sid=85eafd29-f950-4be5-94fd-ac673c3bef37%40sessionmgr4008 >

[100] Ankita, Nabizath Saleenaa [año de publicación] 2018 An Ensemble Classification System for Twitter Sentiment Analysis. Disponible en < https://www-sciencedirect-com.ucatolica.basesdedatosezproxy.com/science/article/pii/S187705091830841X>

[101] Samah Mansour [año de publicacion] 2018 Social Media Analysis of User's Responses to Terrorism Using Sentiment Analysis and Text Mining. Disponible < https://www-sciencedirect-com.ucatolica.basesdedatosezproxy.com/science/article/pii/S1877050918319707 >

[102] Carlos Arcila-Calderón, Félix Ortega-Mohedano, Javier Jiménez-Amores y Sofía Trullenque [año de publicacion] 2017 Supervised sentiment analysis of political messages in spanish: Real-Time of tweets based on machine learning. Disponible en < http://www.elprofesionaldelainformacion.com/contenidos/2017/sep/18.pdf >

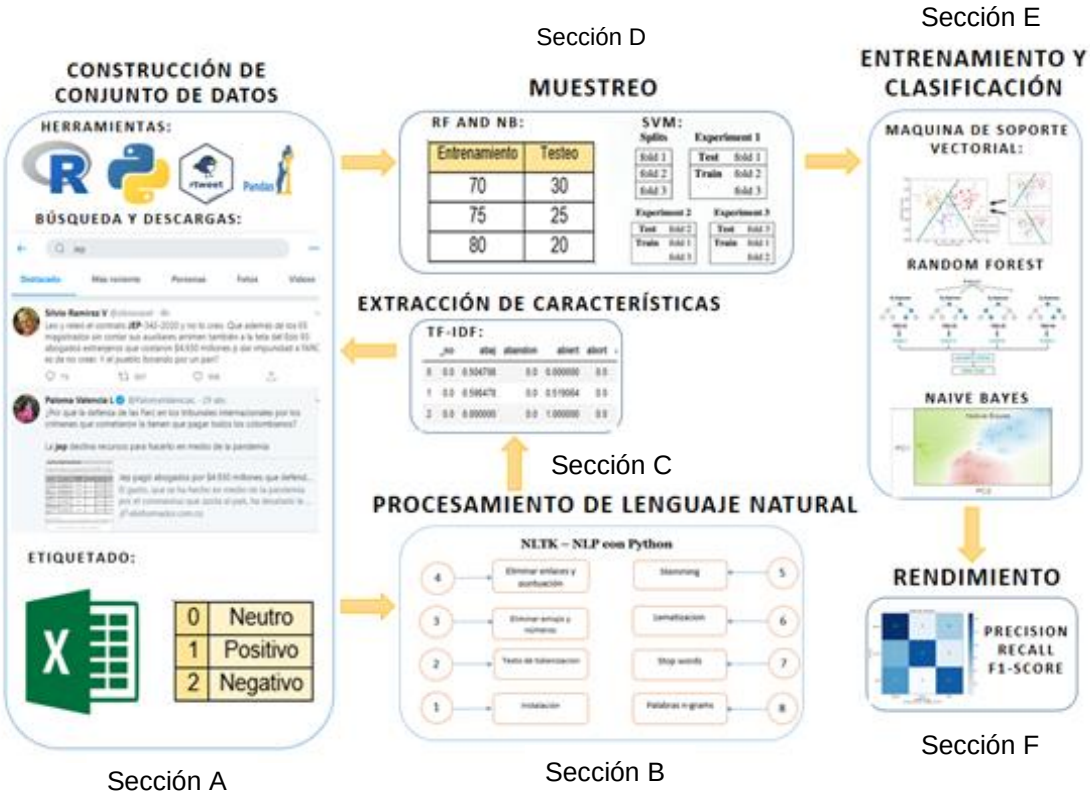
[103] Lina andre torres samboni [año de publicación] 2015 análisis de sentimientos sobre el posconflicto colombiano utilizando herramientas de mineria de text. Disponible en < https://repositorio.escuelaing.edu.co/bitstream/001/403/1/Torres%20Samboni%2C%20Lina%20Andrea%20-%202016.pdf >

El análisis de sentimientos ha sido un campo bastante estudiado en los últimos diez años, por lo general se realiza en el idioma ingles ya que en lo que respecta al español son muy escasos los documentos encontrados. Se descubrió que se realiza análisis de sentimientos para lograr a conocer la tendencia de las personas frente a las elecciones políticas ya que el enfoque más común es conocer si un grupo o población está de acuerdo con determinado tema, en cada documento se aplican diferentes métodos y se comparan los resultados, uno de los casos también encontrados es que se ha realizado un análisis de tendencias al posconflicto en Colombia, pero no se ha tenido en cuenta el tema de la jurisdicción especial para la paz.

3 METODOLOGÍA

Para realizar el proceso de análisis de sentimientos se implementaron una serie de etapas donde cada una dependerá de la anterior, para completar cada etapa se tiene que realizar una serie de tareas específicas, las cuales son extracción de datos, limpieza del conjunto de datos, procesamiento del lenguaje natural, muestreo, entrenamiento y clasificación. Como se logra observar en la Figura 233 que cuenta con 6 etapas para completar el proceso.

Figura 23 Grafica de metodología



Fuente: Los autores.

Conjunto de datos: En la primera etapa como se logra evidenciar en la Figura 233 sección A. Se recopilo la máxima cantidad de tweets, este conjunto de datos se obtiene desde el comienzo de la implementación de la metodología, a partir de la creación de una cuenta de twitter, la cual se debe convertir en una cuenta de desarrollador, con el fin de tener acceso al API de la aplicación. El acceso a esto es necesario para introducir las claves proporcionadas por el API de twitter en un aplicativo cuyo propósito es extraer tweets, ya que de esta manera se logró realizar la creación de un conjunto de datos. El siguiente paso es realizar el etiquetado de los tweets según su sentimiento (positivo, negativo o neutro) como se muestra en las convenciones en la Figura 233.

Procesamiento del lenguaje natural: Es la etapa luego de haber obtenido el primer conjunto de datos como se ve en la Figura 233 sección B, se realiza el procesamiento inicial del texto, en lo que se ejecuta es la tokenización en la cual se divide la oración en palabras para facilitar la limpieza del texto, después se procede a eliminar las stopwords que son las palabras comunes y poco informativas del léxico, el siguiente paso es lematización y Stemming esto consiste en reducir cada palabra a su raíz, eliminando cualquier tipo de derivación. Con el fin de reducir el texto que se quiere procesar y utilizar solo lo más relevante, esta etapa es importante para obtener resultados más exactos.

Extracción de características: Como se ve en la Figura 233 sección C, después de realizar el procesamiento de lenguaje natural, se procede a iniciar la extracción de características que consiste en implementar TF-IDF, el cual representa de manera numérica el texto del conjunto de datos. La extracción de características es convertida en un conjunto de datos, el cual se divide en dos partes, la primera es el conjunto de unigramas y el segundo de bigramas.

Muestreo: Esta etapa de la Figura 233 sección D se separan el conjunto de datos con el fin de obtener la información para el entrenamiento y el testeo de cada algoritmo en el cual se ejecutaron de manera diferente, para el caso de Máquina de Soporte Vectorial se le aplicara KFolds, para el caso de Naive Bayes y Random Forest se dividirán los datos en porcentajes de la siguiente manera 70, 75 y 80% que corresponde al entrenamiento, por otra parte 30, 25 y 20% que es la prueba. Este procedimiento es esencial para realizar el paso de la Sección E.

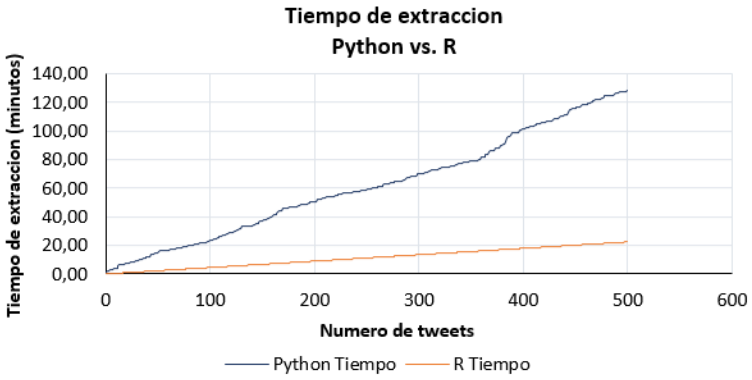
Entrenamiento y clasificación: En la Figura 233 sección E, se ejecuta el entrenamiento de los algoritmos de Máquina de Soporte Vectorial, Random Forest y Naive Bayes. Los cuales realizaron los pasos de entrenamiento y prueba, de acuerdo con los porcentajes mencionados en la sección D.

Desempeño: En esta etapa final en la Figura 23 de la sección F, Se evalúa el porcentaje del desempeño de los algoritmos, el cual se obtiene mediante las métricas de precisión, recall y $\mathrm{F}_1 - \mathrm{Score}$. Estas se lograron obtener con la matriz de confusión, basadas en la clasificación de la sección E.

4 DISEÑO METODOLÓGICO

Conjunto de datos.

En esta etapa se realiza la recolección de los tweets, que tiene como contenido temas relacionados de la JEP, para realizar la descarga de estos se tuvieron en cuenta dos librerías, los cuales son tweepy de Python y rtweet de R. Con el fin de comparar cuál herramienta es más óptima se descargaron los tweets que tenían la palabra JEP en su contenido, estuvieran en idioma español y su antigüedad no fuera mayor a 7 días, al ser un experimento para probar el rendimiento de las herramientas en el momento de descargar tweets, solo se guardaron 500 tweets en formato csv, en donde se midió el tiempo de descarga de cada herramienta, con el fin de conocer cuál de las dos herramientas es más óptima para la realización de la descarga de los datos necesarios, la comparación de estas herramientas se evidencia en la Figura 244.
Figura 24 Tiempo de extracción



Fuente: Los autores.

A lo largo del desarrollo de la metodología se logró obtener una cantidad total de 25000 tweets los cuales son descargados en formato csv, el criterio de búsqueda que se aplicó para obtener esta cantidad fue que los tweets tuvieran la palabra JEP al menos una vez, fueran escritos en español y que su antigüedad no fuera mayor a dos años.

Todos los tweets fueron etiquetados con la ayuda de la herramienta Excel, donde se realizó una limpieza y se eliminaron los tweets repetidos, después de realizar esto, el conjunto de datos final cuenta con 7317 tweets, luego se hizo la clasificación de estos respecto a cada sentimiento expresado mediante cada tweet, estos fueron clasificados como se evidencia en la
. Al finalizar este proceso se obtuvieron 3678 tweets neutros, 2788 negativos y 851 tweets positivos.

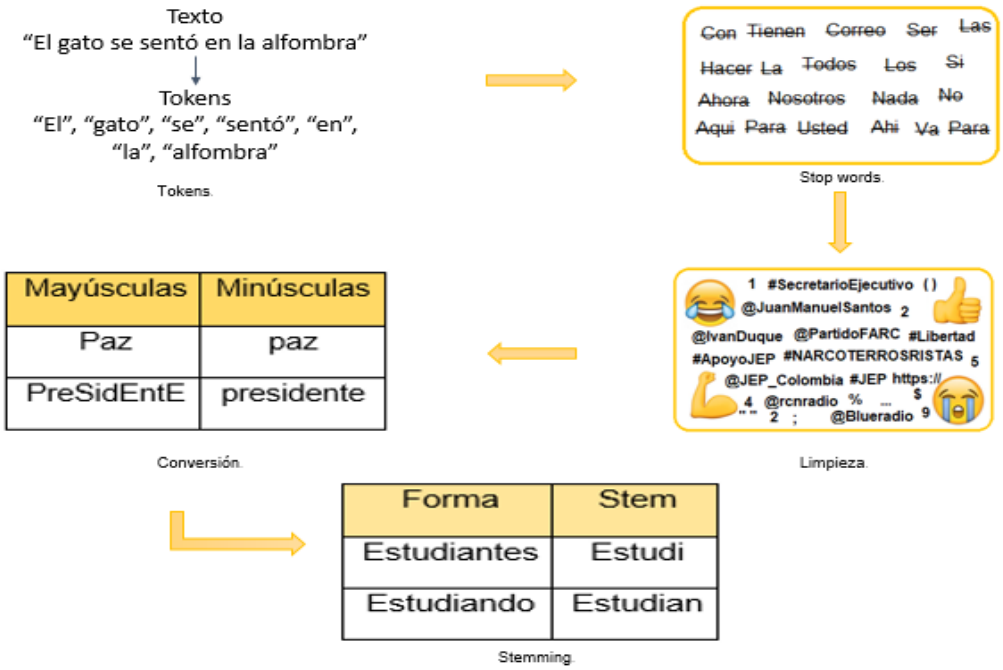
Sentimiento	Etiqueta
Neutral	0
Positivo	1
Negativo	2

Tabla 1 Etiqueta

Procesamiento de lenguaje natural.

Después de realizar el etiquetado se procede a exportar el conjunto de datos a formato csv de nuevo. Utilizando Python se carga el conjunto de datos y se procede a realizar tokenización del texto, mediante un proceso donde cada tweet es separado palabra por palabra, el siguiente paso es hacer una limpieza del texto eliminando hashtags, menciones, signos de puntuación, enlaces de la web, números y emoticones. Lo siguiente es lematizar el texto, el cual consiste en llevar cada palabra a su raíz, en donde relaciona una palabra flexionada o derivada con su forma canónica o lema. Y un lema no es otra cosa que la forma que tienen las palabras cuando las buscas en el diccionario[104], como se muestra en la Figura 255:

Figura 25 Procesamiento lenguaje natural



Fuente: Los autores.

[104] Medium.com [en línea] Lematizacion y stemming <https://medium.com/qu4nt/reducir-el-n%C3%BAmero-de-palabras-de-un-texto-lematizaci%C3%B3n-y-radicalizaci%C3%B3n-stemming-con-python-965bfd0c69fa>

Extracción de características

Después del procesamiento de lenguaje natural se adjunta los tokens, los cuales son separados dependiendo del tipo de n-gram asignado, para el desarrollo de la metodología se utilizó unigramas y bigramas. En la *Tabla 2 N-Gram* se evidencia la cantidad de filas que son representadas por el total de tweets y las columnas por los tokens.

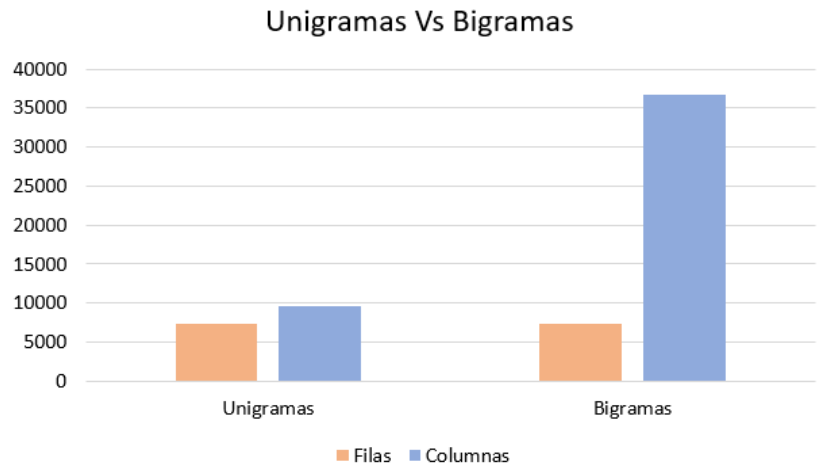
Data sets unigramas vs Bigramas

Tipo	Filas	Columnas	0	1	2
Unigramas	7317	9641	3678	851	2788
Bigramas		36676			

Tabla 2 N-Gram

En la Figura 266, la comparación entre bigramas y unigramas, donde se evidencia que bigramas cuenta con un mayor conjunto de datos puesto que en el momento de realizar n-gram la cantidad de este se incrementa en comparación de los datos que unigramas.

Figura 26 Conjunto de datos Unigramas VS Bigramas



Fuente: Autores.

A su vez se realiza el terminó de frecuencia TF-IDF el cual consiste en asignarle un valor a cada token, este es calculado aplicando las fórmulas de frecuencia de termino y frecuencia inversa de documento. De este procedimiento se obtiene una matriz la cual es evidenciada en la *Figura 277* donde se logra observar un ejemplo de cómo es el resultado de este proceso.

Figura 27 Extracción de características

	_no	abaj	abandon	abiert	abort	abri	abrim	absoluci	absolv	aburri	...
0	0.0	0.504708	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	...
1	0.0	0.598478	0.0	0.519064	0.0	0.0	0.0	0.0	0.0	0.0	...
2	0.0	0.000000	0.0	1.000000	0.0	0.0	0.0	0.0	0.0	0.0	...

Fuente: Los autores.

Muestreo.

Para realizar el muestreo en los algoritmos, los cuales son Máquina de Soporte Vectorial, Random Forest y Naive Bayes, estos se escogieron a partir del estado del arte, puesto que fueron los más utilizados y contaban con los mejores resultados en toda la investigación. El procesamiento que se implementa en el algoritmo de Maquinas de Soporte Vectorial, consiste en realizar K-Folds en 5 grupos o folios donde 4 de estos serán para entrenamiento y uno para realizar la evaluación del modelo, donde cada uno de estos se va iterando, de esta manera se logra aumentar el performance de este algoritmo. Para el muestreo de Random Forest y Naive Bayes, se realizó mediante entr

Tabla 3, a cada algoritmo se le aplica estos porcentajes, ya que de esta manera se logra conocer en qué momento los algoritmos alcanzan un mejor desempeño. Este muestreo se aplica tanto para unigramas, como bigramas.

Entrenamiento	Testeo
70	30
75	25
80	20

Tabla 3 Muestreo

Entrenamiento:

Con el propósito de realizar el entrenamiento se utiliza la librería sklearn una librear muy útil para realizar aprendizaje de máquina, la cual cuenta con algoritmos que se utilizan para entrenar modelos de aprendizaje automático en Python. Para el presente trabajo se utilizan tres algoritmos para realizar aprendizaje de máquina.

El primer algoritmo que se realiza es Naive Bayes, este es un modelo probabilístico de aprendizaje automático, es muy común en la clasificación de texto, está parametrizado por vectores para cada clase (positivo, negativo o neutral) donde el tamaño de cada vector es la cantidad de características, el algoritmo calcula la probabilidad de que cada característica de manera individual aparezca en una muestra perteneciente a una clase.

El segundo algoritmo, Random Forest consiste en crear y combinar aleatoriamente los múltiples árboles de decisión, es un meta estimulador que ajusta distintos clasificadores y utiliza el promedio para mejorar la precisión de predicción, para el clasificador de Random Forest es necesario indicar la cantidad de árboles a implementarse, en el caso de este experimento debido a la extensa cantidad de datos, el propósito es minimizar el tiempo de respuesta por esto la cantidad de árboles que se implementan son 1000.

Por último el algoritmo de máquina de soporte vectorial, este es un modelo de aprendizaje automático supervisado, para este algoritmo se necesita realizar una búsqueda de hyperparametros, según la configuración de estos puede aumentar o disminuir el performance del algoritmo, con el fin de saber cuál es la mejor opción se utilizó la librería GridSearchCV, esta utiliza un diccionario de hyperparametros para calcular el desempeño del algoritmo, haciendo uso de validación cruzada e indicar que configuración obtiene el mejor desempeño, los hyperparametros contemplados se observan en la

C	Kernel	Gamma
0,01	Linear	0,000000001
0,10	Rbf	0,00000001
1,00	Poly	0,0000001
10,00		0,000001
100,00		0,00001
1000,00		0,0001
10000,00		0,001
100000,00		0,01
1000000,00		0,1
10000000,00		1
100000000,00		10
1000000000,00		100
10000000000,00		1000

Tabla 4.

C	Kernel	Gamma
0,01	Linear	0,000000001
0,10	Rbf	0,00000001
1,00	Poly	0,0000001
10,00		0,000001
100,00		0,00001
1000,00		0,0001
10000,00		0,001
100000,00		0,01

1000000,00		0,1
10000000,00		1
100000000,00		10
1000000000,00		100
10000000000,00		1000

Tabla 4 Valores contemplados para C, Gamma y Kernel

Al ejecutar la búsqueda de hiperparámetros utilizando la librería GridSearchCV en Python, se logró encontrar el valor sobresaliente para los hiperparámetros C, gamma y kernel de cada conjunto de datos de características como se evidencia en la Tabla 5.

Tipo	C	Kernel	Gamma
Unigramas	0.1	Linear	1e-3
Bigramas	0.1	Linear	1e-3

Tabla 5 Mejores hiperparametros para cada data set

Los valores de la Tabla 5 se encontraron al ejecutar la búsqueda de hiperparámetros, estos valores de búsqueda se contemplan en los siguientes mapas de calor, con las siguientes Figuras:

Figura 28 Mapa de calor Unigramas kernel Linear

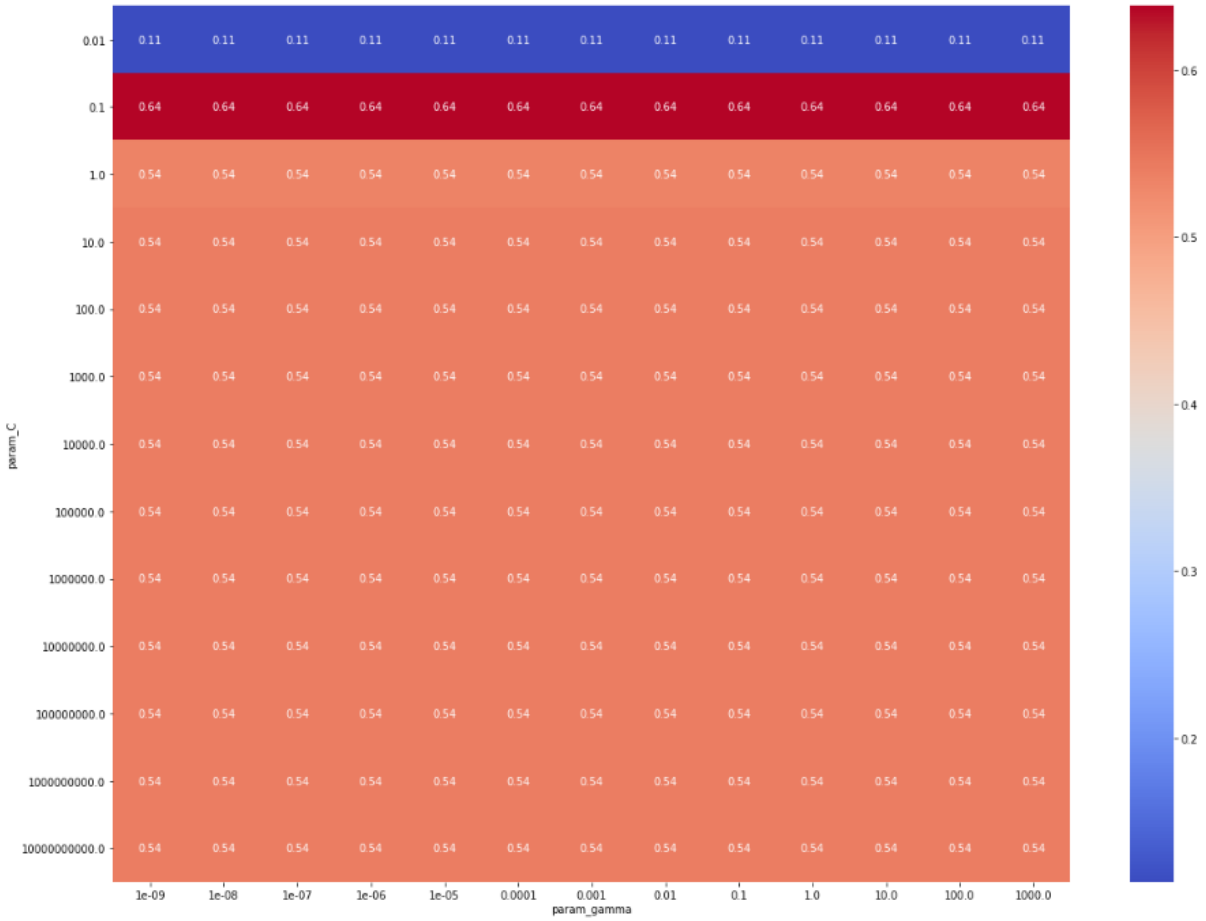


Figura 29 Mapa de calor Unigramas kernel RBF.

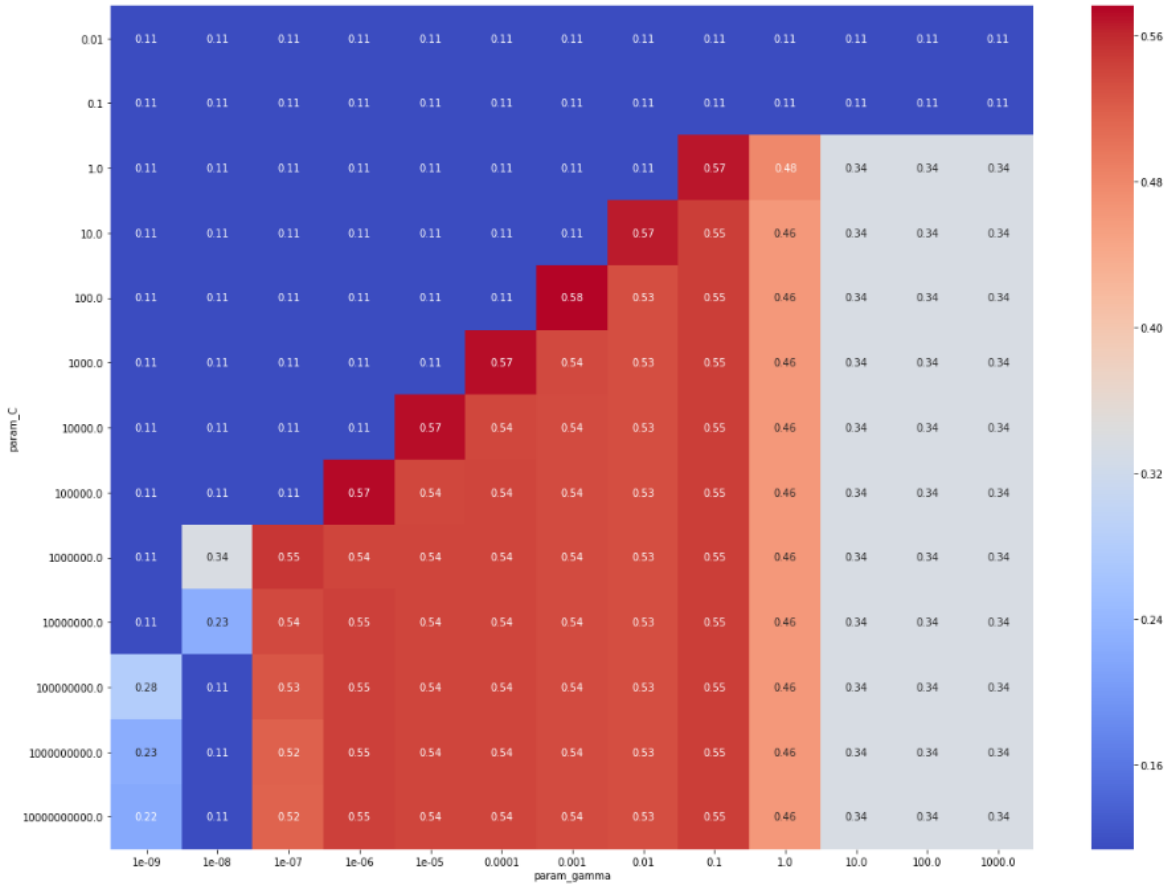


Figura 30 Mapa de calor Unigramas kernel Poly

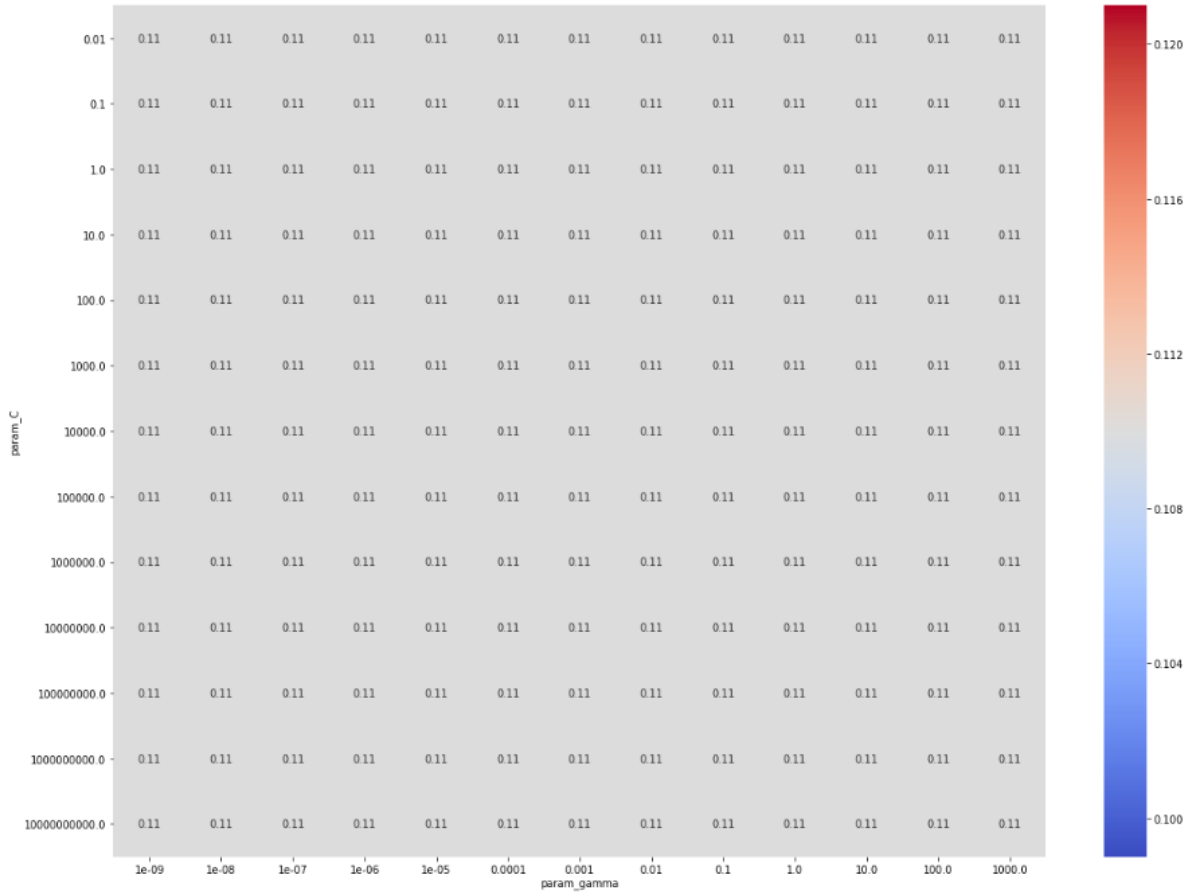


Figura 31 Mapa de calor Bigramas kernel Linear

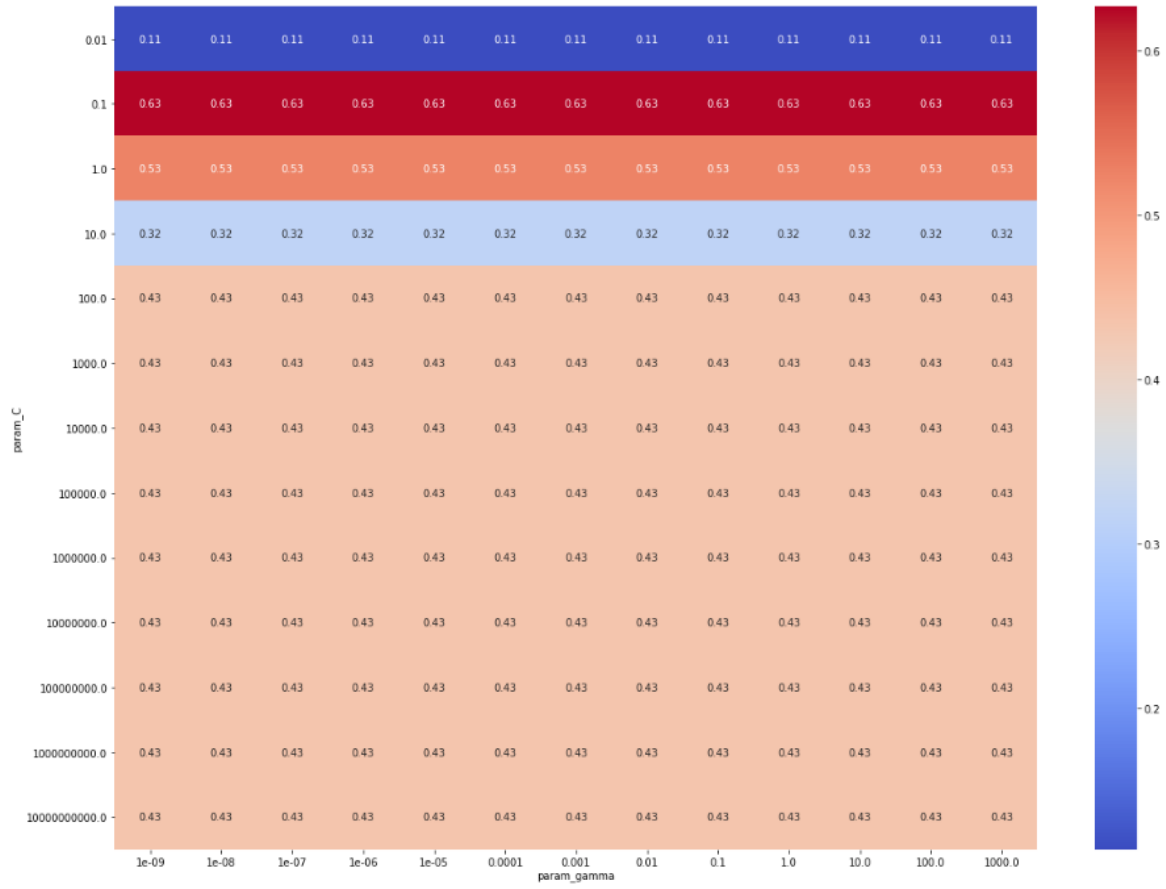


Figura 32 Mapa de calor Bigramas kernel RBF

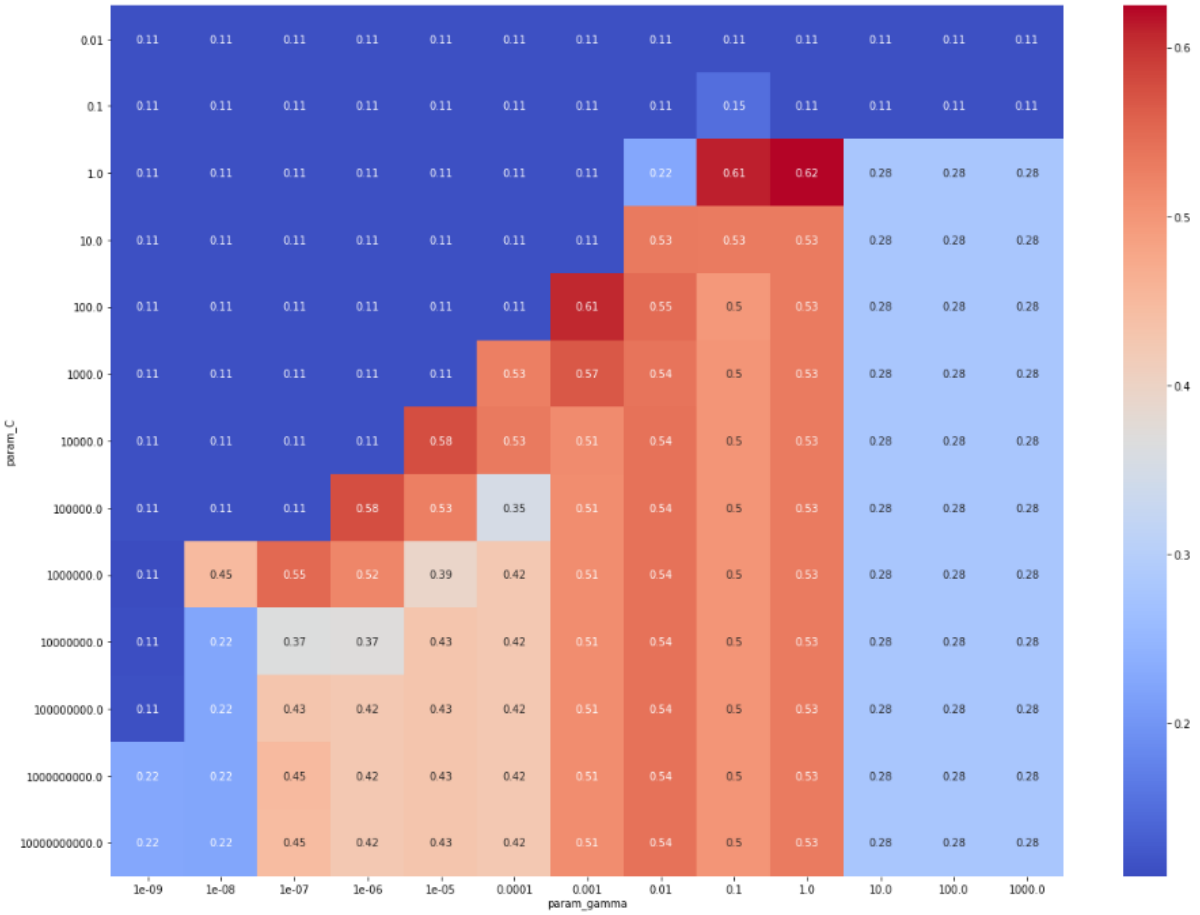
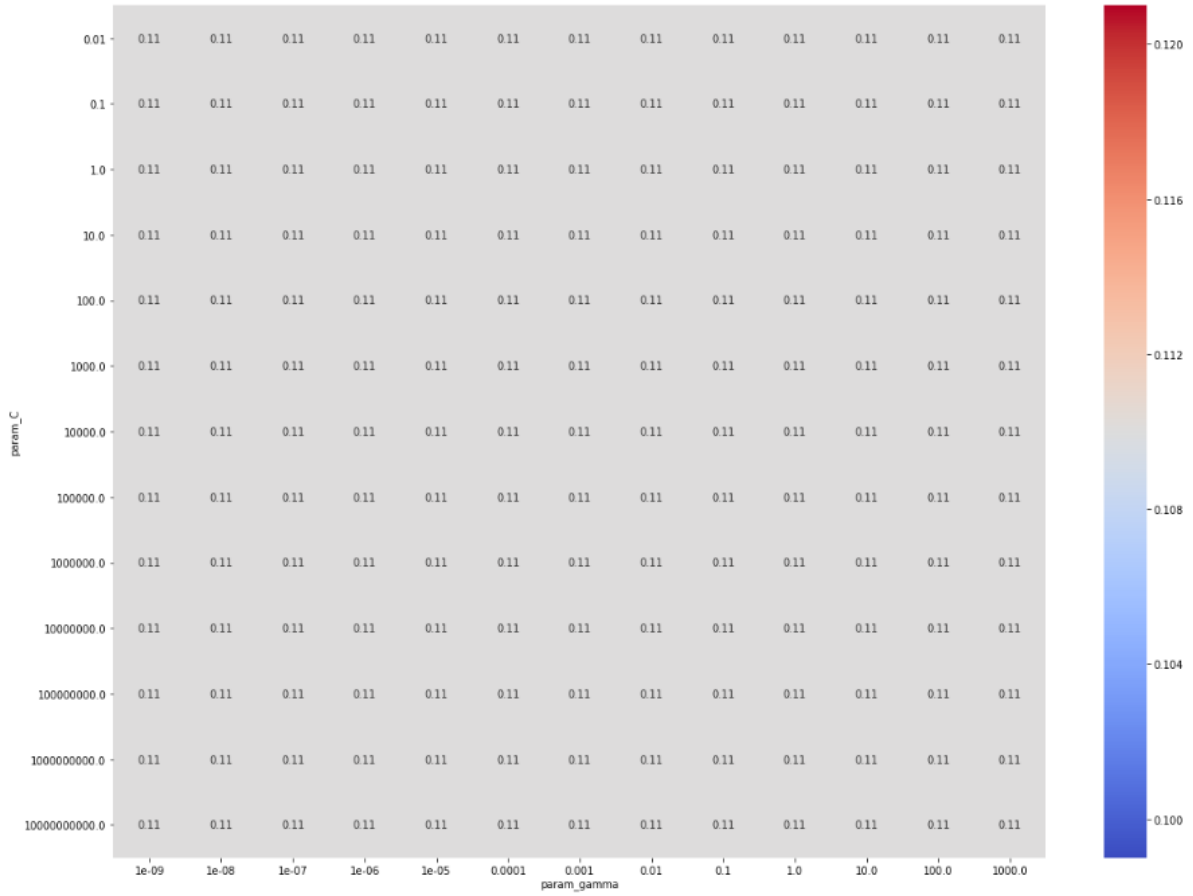


Figura 33 Mapa de calor Bigramas kernel Poly



Clasificación:

Una vez implementado el entrenamiento de cada algoritmo, se procede a ejecutar la clasificación del porcentaje de datos destinado para el testeo del algoritmo Naive Bayes y Random Forest. Por otra parte, con el algoritmo de Maquina de Soporte Vectorial se utilizan todos los datos, ya que en el momento de su clasificación esta efectúa validación cruzada a cada grupo de datos que es seleccionado para este procedimiento.

Al realizar la clasificación el algoritmo nos indica a que clase pertenece cada grupo de datos, esta clase es comparada con la clase asignada al grupo de datos en el momento del etiquetado con el fin de calcular el número de falsos positivos, falsos negativos, verdaderos positivos, verdaderos negativos. La clasificación es una etapa esencial para calcular las medidas de desempeño.

Evaluación del desempeño:

Para medir el rendimiento de los algoritmos se utilizaron las medidas de desempeño precisión, recall, $F_1 - Score$, los resultados obtenidos después del entrenamiento de cada modelo con unigramas y bigramas, se encuentra en las siguientes tablas.

Resultados										
Medida	Precision			Recall			F1-score			Exactitud
clase	0	1	2	0	1	2	0	1	2	
RF - Unigramas	0,66666667	0,8181818	0,7117962	0,8423913	0,1764706	0,634409	0,7442977	0,2903226	0,6708781	0,68579235
	0,73221491			0,551090165			0,56849946			macro pro
	0,701461668			0,68579235			0,663598312			Pro. Ponderado
RF - Bigramas	0,61449647	0,65625	0,6909091	0,8677536	0,0823529	0,499403	0,7194893	0,1463415	0,5797503	0,636156648
	0,653885188			0,483169731			0,481860369			macro pro
	0,648469383			0,636156648			0,599674088			Pro. Ponderado
NB - Unigramas	0,71536524	0,2809917	0,6111111	0,5144928	0,5333333	0,670251	0,5985248	0,368065	0,6393162	0,576047359
	0,535822695			0,572692328			0,535301985			macro pro
	0,625189488			0,576047359			0,58731129			Pro. Ponderado
NB - Bigramas	0,72097054	0,1834061	0,6486486	0,3768116	0,6588235	0,544803	0,4949435	0,2869342	0,5922078	0,473588342
	0,5176751			0,526812664			0,458028507			macro pro
	0,630983129			0,473588342			0,507861459			Pro. Ponderado

Tabla 6 Muestreo 70 – 30

Se observa en la Tabla 6 que se ejecutó un entrenamiento de 70% con un testeo de 30%, con los algoritmos de Random Forest y Naive Bayes con las características de unigramas y bigramas para cada uno, donde los mejores promedios de desempeño para las métricas se dieron en el algoritmo Random Forest unigramas.

Resultados										
Medida	Precision			Recall			F1-Score			Exactitud
clase	0	1	2	0	1	2	0	1	2	
RF - Unigramas	0,677308024	0,79166667	0,72552167	0,85326087	0,17840376	0,64849354	0,7551708	0,2911877	0,6848485	0,696721311
	0,731498787			0,560052723			0,577068993			macro pro
	0,708981959			0,696721311			0,674382228			Pro. Ponderado
RF - Bigramas	0,610687023	0,59259259	0,69371197	0,86956522	0,07511737	0,49067432	0,7174888	0,1333333	0,5747899	0,632786885
	0,632330528			0,478452302			0,475204013			macro pro
	0,640203019			0,632786885			0,595146589			Pro. Ponderado
NB - Unigramas	0,730354391	0,27696078	0,61319534	0,51521739	0,53051643	0,68005739	0,6042065	0,3639291	0,644898	0,579781421
	0,540170173			0,575263737			0,537677869			macro pro
	0,632959476			0,579781421			0,591738124			Pro. Ponderado
NB - Bigramas	0,748945148	0,18134715	0,65068493	0,38586957	0,657277	0,54519369	0,5093257	0,284264	0,5932865	0,478142077
	0,52699241			0,529446749			0,462292045			macro pro
	0,64545567			0,478142077			0,51510849			Pro. Ponderado

Tabla 7 Muestreo 75 – 25

Se visualiza en la Tabla 7 que se implementó un entrenamiento de 75% con un testeo de 25%, con los algoritmos de Random Forest y Naive Bayes con las características de unigramas y bigramas para cada uno, donde los mejores

promedios de desempeño para las métricas se dieron en el algoritmo Random Forest unigramas.

Resultados										
Medida	Precision			Recall			F1-score			Exactitud
clase	0	1	2	0	1	2	0	1	2	
RF - Unigramas	0,6830065	0,8292683	0,7247525	0,8519022	0,2	0,655914	0,758162	0,3222749	0,6886171	0,701502732
	0,745675768			0,569272051			0,589684678			macro pro
	0,715901845			0,701502732			0,681039849			Pro. Ponderado
RF - Bigramas	0,6120114	0,6818182	0,6743003	0,8722826	0,0882353	0,4749104	0,7193277	0,15625	0,5573081	0,629781421
	0,656043292			0,478476099			0,477628609			macro pro
	0,643858642			0,629781421			0,592189637			Pro. Ponderado
NB - Unigramas	0,7235772	0,2735562	0,6111975	0,4836957	0,5294118	0,7043011	0,5798046	0,3607214	0,6544546	0,573087432
	0,536110326			0,572469497			0,531660208			macro pro
	0,628487443			0,573087432			0,582817268			Pro. Ponderado
NB - Bigramas	0,7596685	0,1736434	0,6564551	0,3736413	0,6588235	0,5376344	0,5009107	0,2748466	0,591133	0,469262295
	0,529922354			0,523366414			0,455630126			macro pro
	0,652279625			0,469262295			0,509048123			Pro. Ponderado

Tabla 8 Muestreo 80 – 20

Se evidencia en la Tabla 8 la implementación entrenamiento de 80% con un testeo de 20%, con los algoritmos de Random Forest y Naive Bayes con las características de unigramas y bigramas para cada uno, donde los mejores promedios de desempeño para las métricas se dieron en el algoritmo Random Forest unigramas.

Resultados										
Medida	Precision			Recall			F1-score			Exactitud
clase	0	1	2	0	1	2	0	1	2	
SVM - Unigramas	0,59239453	0,68333333	0,63088768	0,8132137	0,04817861	0,49964132	0,68545892	0,09001098	0,55764612	0,604756048
	0,635538516			0,453677879			0,444372005			macro pro
	0,617638187			0,604756048			0,567505074			Pro. Ponderado
SVM - Bigramas	0,60711091	0,43119266	0,59650516	0,73817292	0,16568743	0,53873745	0,66625767	0,23938879	0,56615153	0,595599289
	0,544936245			0,480865931			0,49059933			macro pro
	0,582609715			0,595599289			0,578467408			Pro. Ponderado

Tabla 9 SVM

Se observa en la Tabla 9 la ejecución del algoritmo de Maquina de Soporte Vectorial con las características de unigramas y bigramas, donde el mejor promedio de desempeño para las métricas se dio con las características de unigramas.

4.2 Instalaciones y equipo requerido

Las instalaciones y el equipo requerido para el desarrollo del proyecto son los que se describen a continuación:

- Las salas de informática presentes en la Universidad católica de Colombia.
- 1 computador portátil con procesador Intel core7, memoria de 6GB de RAM y al menos 200 GB de espacio libre en el disco duro.
- Anaconda con python3.6
- Librerías para el experimento (pandas, numpy, operator, re, string, nltk, sklearn).
- 1 computador portátil con: Procesador Intel Core I5 de decima generación 1.2Ghz, Memoria de 8 GB de RAM, al menos 200 GB de espacio libre en el disco duro.
- Knime
- Computador con acceso a internet y programas necesarios.

4.3 Estrategias de comunicación y divulgación

Como métodos de comunicación y divulgación se tendrán en cuenta lo siguiente:

- La sustentación de proyecto de grado ante los jurados.
- Artículo de investiga

5 RESULTADOS

5.1 Desarrollo de los objetivos:

El primer objetivo, Construir un conjunto de datos de la red social twitter con referencia al tema de la JEP, implementado herramientas que puedan conectarse al API de twitter para descargar su data. Se realizó una comparación mediante dos herramientas las cuales son Python y R para la extracción de los datos, esto se logra evidenciar en la Figura 24 *Tiempo de extracción*, página 63 con esto se demuestra que la herramienta más óptima es R por la cantidad de tweets que logra extraer por minuto. Para la construcción del conjunto de datos se descargaron 25.000 tweets que tenían la palabra JEP en su contenido, en donde se descartando los tweets que se encontraban repetidos mediante los filtros de Excel, de esta manera se clasificaron los tweets como se observa en la Tabla 1. Al finalizar este procedimiento se obtuvieron 7317 tweets.

En el segundo objetivo, diseño de una estrategia de minería de datos para analizar los sentimientos de tweets respecto a la JEP, realizando una investigación con la cual se busca ordenar y ejecutar una serie de pasos para realizar este proceso de manera óptima. Se implementó procesamiento de lenguaje natural esto consistió en realizar tokenizacion, lematizacion, del texto en donde se lleva la palabra a su raíz, separando cada palabra y eliminando hastags, menciones, signos de puntuación, links, números y emoticones, esto se evidencia en la Figura 255, página 64. También se realizó extracción de características, donde se ejecutó n-gram, donde el conjunto de datos fue separado en dos partes que son unigramas y bigramas, su extracción de características nos da una matriz con un numero de filas y columnas para cada n-gram, esto se observa en la Tabla 2, página 65.

En el tercer objetivo, implementación de un modelo de minería de datos para clasificar los tweets en base a los sentimientos de los usuarios para conocer su posición con respecto a la JEP, aplicando algoritmos de aprendizaje de máquina al conjunto de datos. Se ejecutaron los algoritmos de Maquina de Soporte Vectorial para su procesamiento se realizó mediante K-Folds con kernel Lineal, para los algoritmos de Random Forest y Naive Bayes su desarrollo se dio mediante entrenamiento y testeo, esto se evidencia en la Tabla *3*, página 66.

El cuarto objetivo, Evaluar el rendimiento de la técnica basada en minería de datos utilizando las diferentes métricas. Mediante los resultados que se obtuvieron de los algoritmos Random Forest, Naive Bayes y Maquina de Soporte Vectorial, logrando medir el rendimiento de cada uno de estos, con las

medidas de desempeño precisión, recall, $F_1 - Score$ y exactitud. En donde se realizó una comparación de cada uno de estos, logrando evidenciar en la Tabla 8 Muestreo 80 – 20 que el mejor desempeño se dio con el algoritmo Random Forest donde la precisión tuvo un valor sobresaliente de 74.56%, recall con un resultado de 70.15%, $F_1 - Score$ obtuvo un porcentaje óptimo de 68.10% y finalmente la exactitud con un valor de 70.15%.

5.2 Diseño del algoritmo:

Random Forest y Naive Bayes:

```
1. Inicio
2. #definir algoritmo de clasificación
3. algoritmo = MultinomialNB OR RandomForestClassiffier
4. #Entrenar con conjuntos de sampling
5. X = [0.3,0.25,0.2]
6. Para cada i en x Hacer
7.     Si i = 0.3 Entonces
8.       X_train = pd.read_csv("X_train_unigram70.csv)
9.       X_test = pd.read_csv("X_test_unigram70.csv")
10.      y_train = pd.read_csv("y_train_unigram70.csv")
11.      y_test = pd.read_csv("y_test_unigram70.csv"
12.      #Entrenar y calcular medidas de desempeño
13.      entrena(X_train,X_test,y_train,y_test,algoritmo)
14.    Si No
15.      Si i = 0.2 Entonces
16.       X_train = pd.read_csv("X_train_unigram80.csv"
17.       X_test = pd.read_csv("X_test_unigram80.csv")
18.       y_train = pd.read_csv("y_train_unigram80.csv")
19.       y_test = pd.read_csv("y_test_unigram80.csv")
20.       entrena(X_train,X_test,y_train,y_test,algoritmo)
21.      Si No
22.       X_train = pd.read_csv("X_train_unigram75.csv")
23.       X_test = pd.read_csv("X_test_unigram75.csv")
24.       y_train = pd.read_csv("y_train_unigram75.csv")
25.       y_test = pd.read_csv("y_test_unigram75.csv")
26.       entrena(X_train,X_test,y_train,y_test,algoritmo)
27.      Fin Si
28.    Fin Si
29.  Fin
```

Maquinas de soporte vectorial:

```
1.  Inicio
2.  #definir algoritmo de clasificación con parámetros
3.  svclassifier = SVC(C=0.1,
  kernel='linear',gamma=0.001,decision_function_shape='ov
  o')
4.  #Leer dataset
5.  dataset = pd.read_csv("Features unigrama.csv")
6.  #Definir arreglo de caracteristicas y de etiquetas
7.  y = dataset.Label
8.  X = dataset.drop('Label', axis=1)
9.  #Entrenar con validación cruzada
10.  y_pred = cross_val_predict(svclassifier, X, y, cv=5)
11.  FIN
```

6 DISCUSIÓN DE RESULTADOS

Después de implementar la metodología se obtuvieron en las Tabla 66, 7, 8 y 9 con las cuales se evidencia los resultados de los algoritmos y de las métricas que se implementaron para la clasificación de los sentimientos sobre el tema de la Jurisdicción Especial para la Paz.

En la Tabla 6 se contempla los resultados entre un 70% y 30% de entrenamiento y testeo, realizando una comparación entre los algoritmos de Random Forest (RF) y Naive Bayes (NB), con las características de unigramas y bigramas. Donde el mayor porcentaje de precisión se obtuvo del algoritmo Random Forest con las características de unigramas con un porcentaje de precisión general de un promedio ponderado de 73.22%, significando que de cada 100 clasificaciones que se realiza en el modelo 73 de estas son predicciones correctas para el sentimiento de cada tweet.

Para la columna de la métrica de recall se evidencia que el mejor resultado es de 68.57% el cual lo obtuvo RF en unigramas. A su vez, $\mathrm{F}_1 - \mathrm{Score}$ su mejor valor fue de 66.35% con el mismo algoritmo y características de recall. Realizando una comparación en la exactitud entre los dos algoritmos y sus características, se evidencia que el desempeño más alto se dio con el algoritmo Random Forest en unigramas con un valor de 68.57%

En la Tabla 7 se puede observar que con el porcentaje de muestreo entre 75% y 25%, con los algoritmos de Random Forest y Naive Bayes con las características de unigramas y bigramas. Se evidencia que con la métrica de precisión el mejor desempeño se dio con un porcentaje de 73,14%, en el algoritmo RF unigramas con macro promedio. En Recall el resultado óptimo fue de 69,67%, con el mismo algoritmo de precisión, pero con un promedio ponderado. Con $\mathrm{F}_1 - \mathrm{Score}$ el resultado más alto fue de 67,43% en RF unigramas con promedio ponderado. Finalmente, con la medida de desempeño de exactitud el resultado sobresaliente se sigue dando con el algoritmo de Random Forest con las características de unigramas.

Con la **¡Error! No se encuentra el origen de la referencia.**8, mediante los resultados de muestreo entre los porcentajes de 80% y 20% de entrenamiento y testeo, realizando una comparación entre los algoritmos de Random Forest (RF) y Naive Bayes (NB) con las características de unigramas y bigramas. Se observa que con la métrica de precisión el mejor de desempeño fue de 74,56% en RF unigramas con macro promedio.

Para la columna de recall con un promedio ponderado de 70,15% con el algoritmo de Random Forest con las características de unigramas. Mediante la medida de $\mathrm{F}_1 - \mathrm{Score}$, el resultado sobresaliente fue de 68,10% con el mismo

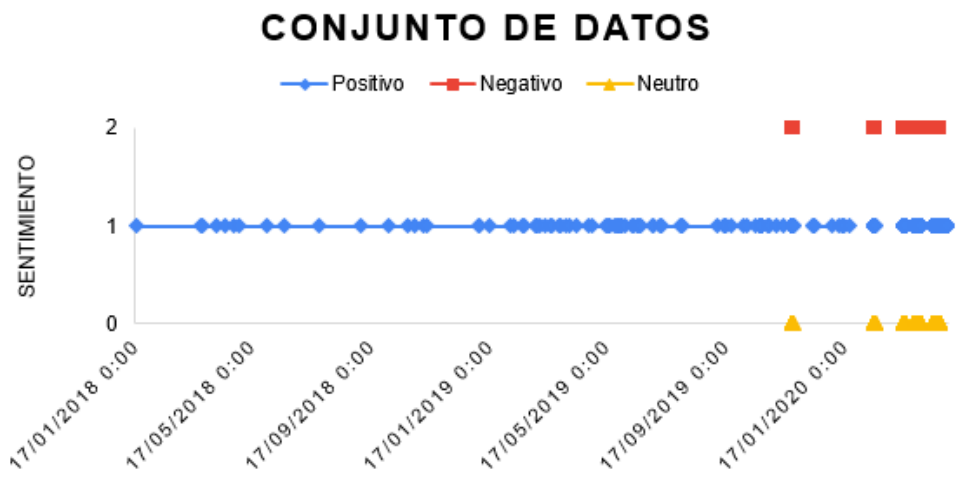
algoritmo mencionado anteriormente. Por último, con la métrica de exactitud para cada algoritmo con sus respectivas características, se evidencia que el mejor resultado fue en Random Forest unigramas con un valor de 70,15%.

Finalmente, en la Tabla 9 los resultados de Máquina de Soporte Vectorial se realizó la búsqueda de hiperparámetros, en donde se ejecutó con el Kernel Lineal ya que su desempeño fue sobresaliente. Al ser ejecutado con las características de unigramas y bigramas, se observa que la precisión tuvo un macro promedio de 63,53% en unigramas. En la métrica de recall el mejor resultado fue con un promedio ponderado de 60.47% con las características de unigramas. Con $\mathrm{F}_1 - \mathrm{Score}$ mediante unigramas obtuvo un promedio ponderado de 56,75%. Por último, con la medida de exactitud el mejor desempeño se dio con un valor de 60,47% en unigramas.

Realizando una comparación con los algoritmos Random Forest, Naive Bayes y Máquina de Soporte Vectorial, con las características de unigramas y bigramas. Se evidencia que la medida más sobresaliente para la métrica de exactitud fue en el algoritmo de Random Forest con las características de unigramas con un entrenamiento y testeo de 80% y 20%, el cual obtuvo un rendimiento de 70.15% es decir que de cada 100 predicciones 70.15 son acertadas, pero al ser un experimento con clases desbalanceadas esta medida puede ser engañosa ya que la mayoría de los tweets son negativos, es muy fácil acertar prediciendo un tweet negativo. A su vez realizando la comparación con $\mathrm{F}_1 - \mathrm{Score}$ el mejor resultado se dio en el algoritmo de Random Forest con unigramas en el mismo muestreo de exactitud, pero su resultado fue de 68.10% lo que significa que el modelo es preciso y tiene un recall bastante alto por lo que de cada 100 predicciones 68.10 van a ser precisas y a su vez recuperadas de una clasificación no acertada. Con los resultados que se obtuvieron, se observa que el mejor desempeño en todas las métricas (precisión, recall, $\mathrm{F}_1 - \mathrm{Score}$ y exactitud) se dio con el algoritmo RF con las características de unigramas, ya que este tiene una predicción mayor para el conjunto de datos que se encuentran desbalanceados como este.

Además, se logra evidenciar que las personas que trinaron durante el tiempo de recolección de datos entre el año 2018 y principios del 2020 como se observa en la Figura 34, la cantidad de trinos positivos fueron disminuyendo a través del tiempo y de esta manera con el conjunto de datos implementado se observa una polarización permitiendo realizar un análisis parar lograr conocer los sentimientos de los usuarios, ya fueran positivos, negativos o neutros.

Figura 34 Conjunto de datos



Fuente: Los autores.

En los mapas de calor de la Figura 288 a la Figura 333 con el conjunto de datos de características unigramas se evidencia que los mejores hiperparámetros se encuentran en el conjunto de kernel linear, ya que al comparar los valores de la zona resaltada en rojo de cada mapa, el performance más alto se obtuvo en la Figura 288 para el C igual a 0.1, dado que los valores de gamma no aumentan o disminuyen el performance del kernel linear a diferencia del kernel RBF donde para cada combinación de C y Gamma el performance da un resultado diferente. Con respecto al kernel poly se obtuvieron los valores de performance más deficientes, esto evidencia que este kernel no es óptimo para el entrenamiento con las características que contienen el conjunto de datos en unigramas, el conjunto de datos de bigramas se evidencia el mismo comportamiento que en el conjunto de datos unigramas, en la figura 30 se observa que el mejor valor para c es 0.1.

7 CONCLUSIONES

- Se logró identificar los sentimientos de los usuarios mediante aprendizaje de máquina, implementando los algoritmos de Random Forest, Naive Bayes y Maquina de Soporte Vectorial midiendo el rendimiento de cada uno con las métricas de $\mathrm{F}_1 - \mathrm{Score}$, exactitud, recall y precisión.

- Con los resultados de la Figura 24 Tiempo de extracción, donde se evidencia la cantidad de tweets y el tiempo de ejecución de las herramientas Python y R, se determinó que la mejor herramienta para la construcción del conjunto de datos es R.

- Se realizó un análisis frente a los experimentos que se han implementado, para la realización e implementación de la metodología, teniendo como base aprendizaje automático junto a los algoritmos de Máquina de Soporte vectorial, Naive Bayes y Random Forest, ejecutando una comparación entre los tres algoritmos para evidenciar cuál de estos es óptimo para la solución de este tipo de problemáticas, ya que son los más implementados.

- Se implementó un modelo de aprendizaje automático junto con minería de datos, en el cual se utilizó procesos de lenguaje natural, TF-IDF, KFolds, kernel, particionamiento y matrices de confusión, para los diferentes algoritmos, con los cuales se clasificaron los sentimientos de los usuarios para lograr analizar su posición frente a la JEP.

- Se evidencio que con la implementación del conjunto de datos que se obtuvo para la creación del proyecto, la cantidad de tweets positivos por parte de los usuarios fueron disminuyendo a través del tiempo.

- Se evaluó el rendimiento de las diferentes métricas, al momento de medir el desempeño de los clasificadores Naive Bayes, Random Forest y Máquina de Soporte Vectorial, se observa que el algoritmo de clasificación Random Forest con las características de unigramas obtuvo un mejor resultado, ya que este tiene una predicción mayor para el conjunto de datos que se encuentran desbalanceados en comparación con los otros clasificadores. De esta manera los resultados de las métricas en este algoritmo obtuvieron un valor de 68.10% en $\mathrm{F}_1 - \mathrm{Score}$, una precisión de 74.56%, un resultado óptimo en recall de 70.15% y una exactitud de 70.15%.

8 RECOMENDACIONES

- Con la evolución de la tecnología, la idea es realizar la comparación de los algoritmos que se pueden implementar en metodologías y problemas como el expuesto en el documento.

- Se recomienda realizar como trabajo a futuro un algoritmo con el cual no solamente se pueda obtener tweets, sino que además los comentarios que se realizan en otras redes sociales como lo es Facebook.

- Es necesario particionar el conjunto de datos que se utilizara dependiendo el algoritmo, porque cada uno de estos tiene un funcionamiento distinto.

- Implementar aprendizaje profundo para identificar cómo cambian los resultados en comparación de aprendizaje automático.

- Realizar autho profiles para identificar si las cuentas donde se twittean son verdaderas o no, puesto que no se realizó ya que su ejecución requiere más tiempo y el diseño de un nuevo estado del arte junto con una nueva metodología.

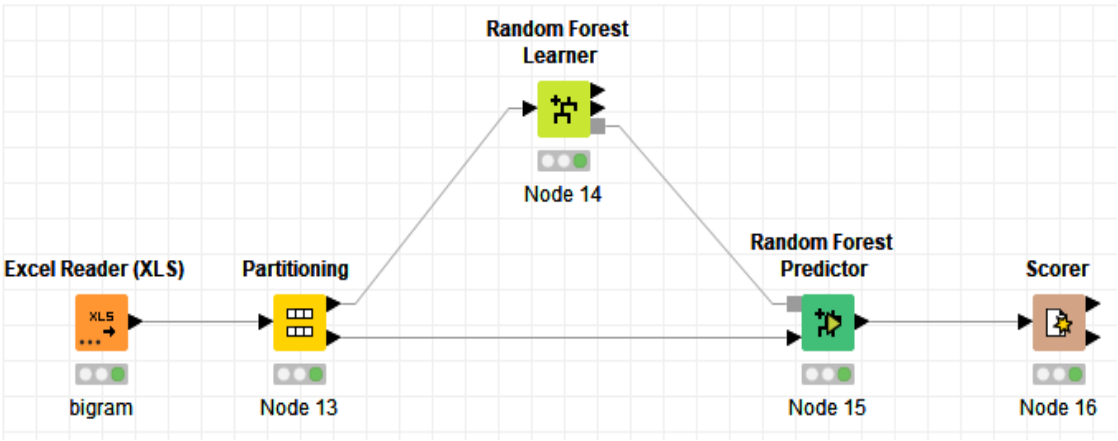
- Implementar el mismo experimento con métricas de Fn-Score.

9 ANEXOS

Anexo A: Resultados con Knime

Para realizar la comparación entre las herramientas de knime analytics platform y Python, la data set que se utilizado para el desarrollo de este anexo es una muestra de la data set real. Para visualizar los resultados mediante Knime el conjunto de datos encuentra separado entre unigramas y bigramas. Después de cargar los datos se realiza el particionamiento de estos, donde son divididos entre aprendizaje y predicción para los algoritmos Naive Bayes (NB) y Random Forest (RF) de manera individual, al finalizar este proceso se logra visualizar los resultados de cada uno de estos, realizando el procedimiento de entrenamiento y testeo por parte de Naive Bayes y Random Forest. En la Figura 35 se observa un ejemplo del diagrama de proceso que se realiza para los algoritmos.

Figura 35 Proceso Knime



Fuente: Los autores.

Para medir el desempeño con esta data set se tuvo en cuenta las métricas de precisión, recall y $F_1 - Score$, donde los resultados que se obtuvieron fueron los siguientes.

Resultados										
Medidas	Precision			Recall			F1-Score			Exactitud
clase	0	1	2	0	1	2	0	1	2	
RF Python- Unigrams	0,381443	0,9375	0,933333	0,986667	0,2	0,186667	0,550186	0,32967	0,311111	0,457778
RF Python - Bigrams	0,511628	0,895833	0,833333	0,88	0,5733333	0,533333	0,647059	0,699187	0,650407	0,662222
NB Python - Unigrams	0,5	0,678571	0,623188	0,48	0,76	0,573333	0,489796	0,716981	0,597222	0,604444
NB Python - Bigrams	0,413333	0,719298	0,611111	0,826667	0,5466667	0,146667	0,551111	0,621212	0,236559	0,506667
RF Knime -Unigrams	0,348	0,889	0,562	0,933	0,107	0,118	0,507	0,19	0,196	0,385
RF Knime - Bigrams	1	0,346	0,467	0,04	0,96	0,092	0,077	0,509	0,154	0,363
NB Knime - Unigrams	0,337	0,255	0,286	0,773	0,16	0,026	0,47	0,197	0,048	0,319
NB Knime - Bigrams	0,322	0,133	0,14	0,907	0,027	0,012	0,486	0,057	0,01	0,31

Tabla 10 Knime - Python 70% y 30%

Resultados										
Medidas	Precision			Recall			F1-score			Exactitud
clase	0	1	2	0	1	2	0	1	2	
RF Python- Unigrams	0,514286	0,782609	0,864865	0,870968	0,571429	0,507937	0,646707	0,6605505	0,64	0,648936
RF Python - Bigrams	0,398693	0,807692	1	0,983871	0,333333	0,142857	0,567442	0,4719101	0,25	0,484043
NB Python- Unigrams	0,440678	0,592105	0,641509	0,419355	0,714286	0,539683	0,429752	0,647482	0,586207	0,558511
NB Python- Bigrams	0,439024	0,730769	0,615385	0,870968	0,603175	0,126984	0,583784	0,6608696	0,210526	0,531915
RF Knime - Unigrams	0,362	1	0,643	0,933	0,107	0,118	0,507	0,19	0,196	0,385
RF Knime - Bigrams	1	0,346	0,467	0,04	0,96	0,092	0,077	0,509	0,154	0,363
NB Knime - Unigrams	0,377	0,255	0,286	0,773	0,16	0,026	0,47	0,197	0,048	0,319
NB Knime - Bigrams	0,322	0,133	0,1	0,907	0,027	0,012	0,476	0,044	0,021	0,31

Tabla 11 Knime - Python 75% - 25%

Resultados										
Medidas	Precision			Recall			F1-Score			Exactitud
clase	0	1	2	0	1	2	0	1	2	
RF Python - Unigrams	0,569444	0,825	0,815789	0,82	0,66	0,62	0,672131	0,733333	0,704545	0,7
RF Python - Bigrams	0,52	0,394958	1	0,26	0,94	0,12	0,346667	0,556213	0,214286	0,44
NB Python - Unigrams	0,45098	0,642857	0,604651	0,46	0,72	0,52	0,455446	0,679245	0,55914	0,5666667
NB Python- Bigrams	0,443299	0,780488	0,5	0,86	0,64	0,12	0,585034	0,703297	0,193548	0,54
RF Knime - Unigrams	0,378	1	0,833	0,96	0,24	0,196	0,542	0,387	0,317	0,464
RF Knime - Bigrams	0,25	1	0,329	0,02	0,02	0,941	0,037	0,039	0,487	0,331
NB Knime- Unigrams	0,354	0,167	0,188	0,7	0,12	0,059	0,47	0,14	0,09	0,291
NB Knime- Bigrams	0,329	0,01	1	0,96	0,01	0,02	0,49	0,16	0,038	0,325

Tabla 12 Knime - Python 80% y 20%

En las Tabla 10, 11 y 12 se realizó una comparación de los resultados que se obtuvieron de las herramientas de Python y Knime, donde se evidencia la implementación de los algoritmos y métricas, para la clasificación de los sentimientos.

Como se evidencia en la Tabla 10 el mejor resultado para la métrica de precisión es de 100% con el algoritmo Random Forest ejecutado con la herramienta knime con las características de bigramas en los tweets neutrales, mientras que para Python el mejor fue en los tweets positivos, el porcentaje de estos es de 93.75% con las características unigramas en RF, por otra parte también se obtuvo un recall alto para RF unigramas en Python el cual tuvo porcentaje de 98.66% en los tweets neutrales, el mayor recall en Knime fue para el mismo algoritmo y características que en Python pero con un valor del 93.3% en los tweets neutrales. Finalmente, el resultado más destacado para la métrica F1-score en Python es de 71.69% en Naive Bayes unigramas en los tweets positivos, con respecto a Knime el mejor resultado es de 50.9% para RF bigramas en los tweets positivos.

En la tabla 11 con un muestreo del 75% y 35%, se puede observar que la mejor precisión obtuvo un porcentaje de 100% en el algoritmo de Random Forest en la herramienta de Python con las características de bigramas en los tweets negativos y en Knime se dio con el mismo algoritmo, pero en las características de unigramas mediante los datos positivos y bigramas con los neutros. Para la

métrica de recall, el mejor desempeño en Python se dio en RF con bigramas el cual dio un valor de 98.38% mediante los tweets neutros, por otra parte, en Knime se dio con el mismo algoritmo y los datos mencionados anteriormente, pero con las características de unigramas que obtuvo un resultado de 93.3%. Para $\mathrm{F}_1 - \mathrm{Score}$ con Python el resultado sobresaliente se dio en NB bigramas con los tweets positivos ya que dieron un valor de 66.08% y Knime se dio en RF en unigramas con datos neutros los cuales tuvieron un porcentaje de 50.7%. Finalmente, el mejor desempeño de exactitud obtuvo un valor de 64.89% para el algoritmo de Random Forest con características de unigramas y para la herramienta de Knime se dio con el mismo algoritmo mencionado anteriormente, pero su resultado fue de 38.5%.

En la Tabla 12 con un muestreo de 80% y 20%, se evidencia que la precisión tuvo un porcentaje de 100% de mediante la herramienta Python con el algoritmo de Random Forest bigramas con los tweets negativos y con Knime se obtuvo con los dos algoritmos RF unigramas y bigramas, pero NB se dio con las características de bigramas. Mediante la métrica recall Python dio un resultado de 94% con el algoritmo de RF sentimientos positivos con bigramas, por otra parte, Knime logro un desempeño de 96% con el mismo algoritmo que en Python, pero con los tweets neutros. Con la métrica de $\mathrm{F}_1 - \mathrm{Score}$ mediante Python alcanzo un valor de 73.33% con los datos positivos en Random Forest unigramas y Knime su mejor desempeñó dio un resultado de 54.2%, que se dio en RF unigramas con los tweets neutros. Por último, la exactitud en Python dio un valor de 56.6% en NB unigramas y en la herramienta de Knime su mejor porcentaje es 46.4% con RF unigramas.

Al comparar las tres tablas se puede observar que los mejores resultados con respecto a la precisión se dieron en knime donde la mayoría de medidas alcanzaron el 100%, con respecto al recall el resultado más sobresaliente se dio en el muestreo 70% y 30% en el algoritmo RF con características unigramas en los tweets neutrales con un porcentaje de 98.66%, con la métrica de exactitud se logró el mejor porcentaje con un valor de 64,89% en RF con unigramas con Python. Finalmente se evidencio que para la métrica $\mathrm{F}_1 - \mathrm{Score}$ el mejor resultado se dio en el muestreo 80% y 20 % con un porcentaje de 73.33% para Random Forest unigramas en los tweets positivos en Python.

Mediante estos resultados podemos evidenciar que la mejor herramienta es Python para la mayoría de métricas. Además, entre más grande sea el muestreo mejor será el resultado de $\mathrm{F}_1 - \mathrm{Score}$ mediante un conjunto de datos, como lo es el que implementamos con una cantidad de 750 filas además Python es mejor al momento de procesar grandes cantidades de datos.

Como se evidencia en el anexo, El conjunto de datos implementado al ser de prueba contiene una menor cantidad, por esta razón el rendimiento de la

métrica $F_1 - Score$ es pequeño, lo que comprueba que a mayor cantidad de datos el desempeño de performance mejora.

Anexo B: Conjunto de datos

Se realiza la entrega del conjunto de datos con el cual se trabajó para el desarrollo del proyecto.

URL: https://kaggle.com/tesisjep/tweets-acerca-de-la-jep-etiquetados

Anexo C: Repositorio de código

Se entrega los códigos ejecutados en las etapas de procesamiento de lenguaje natural, extracción de características, muestreo, entrenamiento y clasificación, rendimiento.

URL: https://github.com/jepresearch/Desarrollo

Anexo D: Repositorio Knime

Se realiza la evidencia de los experimentos ejecutados en la herramienta de Knime.

URL: https://github.com/jepresearch/Desarrollo-knime

10 BIBLIOGRAFÍA

jep.gov.co [en línea] ¿Que es la jurisdicción especial para la paz? <https://www.jep.gov.co/Infografas/conozcalajep.pdf>

elmundo.com [en línea] Acuerdo de paz: errores en serie < https://www.elmundo.com/noticia/Acuerdo-de-pazerrores-en-serie/377466>

eltiempo.com [en línea] El 11% de la población urbana en Colombia usa Twitter Disponible en internet <https://www.eltiempo.com/archivo/documento/CMS-12406882>.

eltiempo.com [en línea] Lo mejor del 2018 en Twitter: las cuentas y los hashtags más populares <https://www.eltiempo.com/tecnosfera/novedades-tecnologia/las-cuentas-y-las-etiquetas-mas-populares-en-twitter-durante-2018-301996>.

diarioinformacion.com [en línea] El éxito de Obama y la minería de datos <https://www.diarioinformacion.com/opinion/2012/11/17/exito-obama-mineria-datos/1315856.html>

bbc.com [en línea] Elecciones en Estados Unidos ¿Fue facebook la clave para el triunfo de Donald Trump? <https://www.bbc.com/mundo/noticias-internacional-37946548>

javeriana.edu.co [en línea] <https://repository.javeriana.edu.co/bitstream/handle/10554/20516/CaicedoOrtizLuisEduardo2016.pdf?sequence=1&isAllowed=y>

elespectador.com [en línea] El 47% de los colombianos tiene una opinión favorable de la jep <https://www.elespectador.com/noticias/politica/47-de-colombianos-tienen-una-opinion-favorable-de-la-jep-gallup-poll-articulo-861085>

lafm.com [en línea] Las polémicas cuentas de la jep <https://www.lafm.com.co/judicial/las-polemicas-cuentas-de-la-jep>

portafolio.com [en línea] El plebiscito por la paz cuesta $350.000 millones, ¿qué se puede hacer con ese mismo dinero? <https://www.portafolio.co/tendencias/cuanto-cuesta-el-plebiscito-por-la-paz-499066>

portafolio.com [en línea] El plebiscito por la paz cuesta $350.000 millones, ¿qué se puede hacer con ese mismo dinero? <https://www.portafolio.co/tendencias/cuanto-cuesta-el-plebiscito-por-la-paz-499066>

mintic.gov.co [en línea] Colombia es uno de los países con más usuarios en redes sociales en la región<https://mintic.gov.co/portal/604/w3-article-2713.html?_noredirect=1>

bbc.com [en línea] Elecciones en Estados Unidos ¿Fue facebook la clave para el triunfo de Donald Trump? <https://www.bbc.com/mundo/noticias-internacional-37946548>

openaccess.uoc.edu [en línea] Análisis de sentimientos en twitter <http://openaccess.uoc.edu/webapps/o2/bitstream/10609/81435/6/jsobrinosTFM0618memoria.pdf>

repository.javeriana.edu.co [en línea] análisis del sentimiento político mediante la aplicación de herramientas de minería de datos a través del uso de redes SOCIALES<https://repository.javeriana.edu.co/bitstream/handle/10554/20516/CaicedoOrtizLuisEduardo2016.pdf?sequence=1&isAllowed=y>

las2orillas.co/ [en línea] Aumenta la polarización política tras la decisión de la JEP sobre Santrich <https://www.las2orillas.co/polarizacion-politica-decision-jep/>

elespectador.com [en línea] 47 % de colombianos tienen una opinión favorable de la JEP: Gallup Poll <https://www.elespectador.com/noticias/politica/47-de-colombianos-tienen-una-opinion-favorable-de-la-jep-gallup-poll-articulo-861085>

bigdata-social.com [en línea] Análisis predictivo<http://www.bigdata-social.com/que-es-el-analisis-predictivo/ >

Christopher M. Bishop. Pattern recognition and machine learning: Linear models for regression En: Information Science and Statics. Pages 45 – 46.

docs.microsoft.com [en línea] Algoritmos de minería de datos<https://docs.microsoft.com/es-es/analysis-services/data-mining/data-mining-algorithms-analysis-services-data-mining?view=sql-server-2017 >

docs.microsfot.com [en línea] Naive Bayes Algorithm <https://docs.microsoft.com/es-es/analysis-services/data-mining/microsoft-naive-bayes-algorithm?view=sql-server-2017 >

Christopher M. Bishop. Pattern recognition and machine learning: Linear models for regression En: Information Science and Statics. Contents 3. Pages 149 – 150.

dataprix.com [en línea] análisis discriminante < https://www.dataprix.com/blog-it/mineria-datos/data-mining-analisis-discriminante-caso-sas>

Christopher M. Bishop. Pattern recognition and machine learning: Linear models for regression En: Information Science and Statics. Contents 2. Pages 90 – 97.

docs.microsfot.com [en línea] Arboles de decisión <https://docs.microsoft.com/es-es/analysis-services/data-mining/microsoft-decision-trees-algorithm?view=sql-server-2017 >

towarddatascience.com [en línea] Arboles de decisión < https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052>

docs.microsfot.com [en línea] Neutral network algorithm <https://docs.microsoft.com/es-es/analysis-services/data-mining/microsoft-neural-network-algorithm?view=sql-server-2017 >

Christopher M. Bishop. Pattern recognition and machine learning: Linear models for regression En: Information Science and Statics. Contents 2. Pages 110.

Tamps.cinvestav.mx [en línea] Minería de datos descriptiva <https://www.tamps.cinvestav.mx/~hmarin/Mineria/EC2.pdf >

Techdifferences.com [en línea] Minería de datos descriptiva< https://techdifferences.com/difference-between-descriptive-and-predictive-data-mining.html >

cs.us.es [en línea] técnicas de clustering < https://www.cs.us.es/~fran/curso_unia/clustering.html>

es.coursea.org [en línea] que es clustering <https://es.coursera.org/lecture/mineria-de-datos-introduccion/que-es-clustering-TMSYv>

educba.com [en línea] Clustering < https://www.educba.com/what-is-clustering-in-data-mining/ >

docs.microsfot.com [en línea] Algoritmos de minería de datos <https://docs.microsoft.com/es-es/analysis-services/data-mining/data-mining-algorithms-analysis-services-data-mining?view=sql-server-2017>

ibm.com [en línea] Segmentation <https://www.ibm.com/support/knowledgecenter/en/SSEPGG_9.7.0/com.ibm.datatools.datamining.doc/miningplan_custseg.html>

stateofdigital.com [en línea] Segmentation < https://www.stateofdigital.com/one-size-does-not-fit-all-data-segmentation/>

docs.microsfot.com [en línea] Algoritmos de minería de datos <https://docs.microsoft.com/es-es/analysis-services/data-mining/data-mining-algorithms-analysis-services-data-mining?view=sql-server-2017>

docs.oracle.com [en línea] association < https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/market_basket.htm#DMCON009>

urg.es [en línea] analisis exploratorio < https://www.ugr.es/~batanero/pages/ARTICULOS/anaexplora.pdf >

stat.cmu.edu [en línea] análisis exploratorio < https://www.stat.cmu.edu/~hseltman/309/Book/chapter4.pdf>

Geeksforgeeks.org [en línea] análisis discriminante lineal < https://www.geeksforgeeks.org/ml-linear-discriminant-analysis/>

Machinelearningmastery.com [en línea] análisis discriminante lineal <https://machinelearningmastery.com/linear-discriminant-analysis-for-machine-learning/>

Uc-r.github.io [en línea] discriminant analysis http://uc-r.github.io/discriminant_analysis

Datascienceblog.net [en linea] linear and quadratic discriminant analysis https://www.datascienceblog.net/post/machine-learning/linear-discriminant-analysis/

docs.microsfot.com [en línea] Conceptos de minería de datos <https://docs.microsoft.com/es-es/analysis-services/data-mining/data-mining-concepts?view=sql-server-2017>

docs.oracle.com [en línea] Conceptos de minería de datos < https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm#DMCON046 >

ibm.com [en línea] Conceptos de minería de texto < https://www.ibm.com/support/knowledgecenter/es/SS3RA7_18.1.1/ta_guide_ddita/textmining/shared_entities/tm_intro_tm_defined.html >

Docs.oracle.com [en línea] About text mining < https://docs.oracle.com/database/121/DMPRG/GUID-3E60BDD1-DE22-494F-8B6D-C73A03EDD01B.htm#DMPRG778 >

Docs.oracle.com [en línea] Text mining <https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/text.htm#BCEDHEDD >

Christopher M. Bishop. Pattern recognition and machine learning: Linear models for regression En: Information Science and Statics. Contents 5. Pages 225 - 236.

arimetrics.com [en línea] Análisis de sentimientos <https://www.arimetrics.com/glosario-digital/analisis-de-sentimiento>

towardsdatascience.com [en línea] Sentiment analysis: concept, analysis and applications < https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17>

lexalytics.com [en línea] Sentiment analysis explained <https://www.lexalytics.com/technology/sentiment-analysis>

cleverdata.io [en línea] conjunto de datos <https://cleverdata.io/conceptos-basicos-machine-learning/>

Ibm.com [en línea] what is a data set? < https://www.ibm.com/support/knowledgecenter/zosbasics/com.ibm.zos.zconcepts/zconc_datasetintro.htm>

docs.microsfot.com [en línea] Modelo de minería de datos <https://docs.microsoft.com/es-es/analysis-services/data-mining/mining-models-analysis-services-data-mining?view=sql-server-2017#bkmk_mdlDefine>

docs.oracle.com [en línea] Predictive analysis <https://docs.oracle.com/cd/E28280_01/admin.1111/e14568/predict.htm#AAMAD5159>

cleardata.io [en línea] Conceptos básicos de Machine Learning <https://cleverdata.io/conceptos-basicos-machine-learning/>

Docs.oracle.com [en línea] supervised data mining < https://docs.oracle.com/cd/B19306_01/datamine.102/b14339/3predictive.htm#i1005885>

jep.gov.co [en línea] Jurisdicción especial para la paz <https://www.jep.gov.co/Paginas/JEP/Jurisdiccion-Especial-para-la-Paz.aspx>

Ema Kušena, Mark Strembeck [año de publicacion] 2018 Politics, sentiments, and misinformation: An analysis of the Twitter discussion on the 2016 Austrian Presidential Elections. Disponible en < https://www-sciencedirect-com.ucatolica.basesdedatosezproxy.com/science/article/pii/S2468696417301088>

Efthymios Kouloumpis, Theresa Wilson, Johanna Moore [año de publicacion] 2010 Twitter Sentiment Analysis:The Good the Bad and the OMG!. Disponible en <

https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2857/3251 >

Trevor Hastie, Robert Tibshirani, Jerome Friedman. The elements of statistical learning, The second edition. En: chapter 12. Page 417 – 418.

Trevor Hastie, Robert Tibshirani, Jerome Friedman. The elements of statistical learning, the second edition. En: chapter 15. Page 587– 589.

Trevor Hastie, Robert Tibshirani, Jerome Friedman. The elements of statistical learning, the second edition. En: chapter 6. Page 210 – 211

Trevor Hastie, Robert Tibshirani, Jerome Friedman. The elements of statistical learning, the second edition. En: chapter 6. Page 191 – 200.

Eric S.Tellez, Sabino Miranda-Jiménez, Mario Graff, Daniela Moctezuma, Oscar S.Siordia, Elio A.Villaseñor [año de publicacion] 2017 A case study of Spanish text transformations for twitter sentiment analysis. Disponible en < https://www-sciencedirect-com.ucatolica.basesdedatosezproxy.com/science/article/pii/S0957417417302312>

Marcela Mayumi Mauricio Yagui, Luís Fernando Monsores Passos Maia, Jonice Oliveira, Adriana S. Vivacqua [año de publicacion] 2018 Data mining of social manifestations in Twitter: Analysis and aspects of the social movement "Bela, recatada e do lar" (Beautiful, demure and housewife) Disponible en < http://web.a.ebscohost.com.ucatolica.basesdedatosezproxy.com/ehost/pdfviewer/pdfviewer?vid=1&sid=85eafd29-f950-4be5-94fd-ac673c3bef37%40sessionmgr4008 >

Ankita, Nabizath Saleenaa [año de publicación] 2018 An Ensemble Classification System for Twitter Sentiment Analysis. Disponible en < https://www-sciencedirect-com.ucatolica.basesdedatosezproxy.com/science/article/pii/S187705091830841X>

Samah Mansour [año de publicacion] 2018 Social Media Analysis of User's Responses to Terrorism Using Sentiment Analysis and Text Mining. Disponible < https://www-sciencedirect-com.ucatolica.basesdedatosezproxy.com/science/article/pii/S1877050918319707 >

Carlos Arcila-Calderón, Félix Ortega-Mohedano, Javier Jiménez-Amores y Sofía Trullenque [año de publicacion] 2017 Supervised sentiment analysis of political messages in spanish: Real-Time of tweets based on machine learning. Disponible en < http://www.elprofesionaldelainformacion.com/contenidos/2017/sep/18.pdf >

Lina andre torres samboni [año de publicación] 2015 análisis de sentimientos sobre el posconflicto colombiano utilizando herramientas de mineria de text.

Disponible en < https://repositorio.escuelaing.edu.co/bitstream/001/403/1/Torres%20Samboni%2C%20Lina%20Andrea%20-%202016.pdf >

Towardsdatascience.com [en línea] Natural language processing feature <https://towardsdatascience.com/natural-language-processing-feature-engineering-using-tf-idf-e8b9d00e7e76>

web.standfor.edu [en linea] The elements of statistical learning <https://web.stanford.edu/~hastie/Papers/ESLII.pdf>

Monkeylearn.com [En línea] Support vector Machine <https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/>

Medium.com [En línea] SVM <https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72>

Towards data science [en línea] Naive Bayes < https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c>

scikit-learn.org [en línea] random forest <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>

Towards data science [en línea] Naive bayes < https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c>

Towards data science [en línea] Cross validation < https://towardsdatascience.com/why-and-how-to-cross-validate-a-model-d6424b45261f>

Towardsdatascience.com [en línea] KFOLDS <https://towardsdatascience.com/natural-language-processing-feature-engineering-using-tf-idf-e8b9d00e7e76>

Towardsdatascience.com [en línea] NLP <https://towardsdatascience.com/natural-language-processing-feature-engineering-using-tf-idf-e8b9d00e7e76>

Towardsdatascience.com [en línea] NGRAM < https://towardsdatascience.com/introduction-to-language-models-n-gram-e323081503d9>

Knime.com [en línea] <https://www.knime.com/about>

Semanticscholar [en línea] <https://pdfs.semanticscholar.org/1d10/6a2730801b6210a67f7622e4d192bb309303.pdf>

Analytics Lane [en línea] <https://www.analyticslane.com/2019/12/16/cual-es-la-diferencia-entre-parametro-e-hiperparametro/>

Investopedia [en línea] https://www.investopedia.com/terms/w/weightedaverage.asp

Machine learning [en línea] <https://developers.google.com/machine-learning/crash-course/classification/accuracy>

Investopedia [en línea] https://www.investopedia.com/terms/w/weightedaverage.asp

Confusion matrix [en línea] <https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62>

SVM with kernel [en línea] <https://towardsdatascience.com/support-vector-machines-svm-c9ef22815589>